UNIVERZITET SINGIDUNUM

DEPARTMAN ZA POSTDIPLOMSKE STUDIJE

Internet pretraživači i inteligentni agenti za pretraživanje Web-a

- Master rad -

Mentor:

Prof. dr Milan Milosavljević

Student:

Aleksić Aljoša

Br. indeksa: 410330/2010

Beograd, 2012.

DEPARTMAN ZA POSTDIPLOMSKE STUDIJE

UNIVERZITET SINGIDUNUM

DEPARTMAN ZA POSTDIPLOMSKE STUDIJE

Beograd, Danijelova 32

Kandidat: Aleksić Aljoša

Broj indeksa: 410330/2010

Smer: Savremene informacione tehnologije

Tema: Ineternet pretraživači i inteligentni agenti za pretraživanje Web-a

Datum odobrenja rada:
Beograd, ___.___._____.

MENTOR

_____

Prof. dr Milan Milosavljević

DEKAN

_____

Prof. dr Mladen Veinović

SADRŽAJ

1. Uvod ..... - 3 -

2. Metodologija Naučnog Istraživanja ..... - 5 -

2.1. Predmet istraživanja ..... - 5 -

2.2. Hipotetički okvir istraživanja ..... - 5 -

2.3. Ciljevi i zadaci istraživanja ..... - 5 -

2.4. Metode i tok istraživanja rada ..... - 6 -

3. Internet pretraživači – metode i analiza ..... - 7 -

3.1. Kako rade internet pretraživači ..... - 7 -

3.2. Internacionalno tržište ..... - 9 -

3.3. Razvoj internet pretraživača ..... - 10 -

3.4. Optimizacija internet pretrage ..... - 12 -

3.5. Uticaj Web-mastera na rad internet pretraživača ..... - 13 -

3.6. Dodatne činjenice o internet pretraživačima ..... - 14 -

3.7. Problemi internet pretraživanja ..... - 17 -

4. Inteligentni agenti, opšti pojmovi, arhitektura ..... - 18 -

4.1. Agenti ..... - 19 -

4.2. Definicije agenata ..... - 20 -

4.3. Podela agenata ..... - 21 -

4.4. Inteligentni agenti ..... - 22 -

4.5. Arhitektura agenata ..... - 23 -

5. Agenti za pretraživanje Web-a ..... - 25 -

5.1. Inteligentni agenti za pretraživanje Web-a ..... - 27 -

5.2. Izgradnja jednostavnog inteligentnog agenta za pretraživanje Web-a ... - 29 -

6. Metode pretraživanja web-a ..... - 32 -

6.1. KQML ..... - 34 -

7. Google ..... - 35 -

7.1. Pojam „googlati“ ..... - 37 -

7.2. Arhitektura ..... - 38 -

7.3. Redosled stranica ..... - 39 -

7.4. Veličina indexa i dodatne činjenice o Google-u ..... - 40 -

7.5. Upotreba ..... - 41 -

7.6. Proširena upotreba ..... - 42 -

7.7. Google Chrome ..... - 43 -

7.8. Podaci koji se čuvaju na računaru prilikom korišćenja Google Chrome - 45 -

7.9. Informacije o korišćenju dodataka za Google Chrome pretraživač ..... - 46 -

7.10. GoogleBot ..... - 47 -

8. Zaključak ..... - 48 -

9. Literatura ..... - 50 -

1. Uvod

Internet pretrazivač je alatka dizajnirana da traži informacije na WWW (World Wide Web) i to uglavnom zadavanjem ključnih reči, a mnogo ređe odabirom ponuđenih stavki. Rezultati pretrage su obično predstavljeni u obliku liste i najčesće se nazivaju " hits". Informacije se mogu sastojati od web stranica, slika, informacija i drugih tipova fajlova. Neki internet pretraživači takodje pretražuju podatke koji se nalaze u "newsbook", bazama podataka ili "open directories". Za razliku od web direktorijuma, koji se održavaju od strane ljudskih editora, internet pretraživači rade na principu algoritma ili su mešavina algoritamskog i ljudskog faktora.

Internet pretraživači su, u osnovi, ekspertski sitemi koji imaju za cilj stvaranje što više heuristika sposobnih za pomoć ekspertskom sistemu u predviđanju šta je to što korisnik traži. Specijalizovni internet pretraživači za svrhu imaju traženje informacija u vezi sa specifičnom oblašću.

Na World Wide Web-u možemo pronaći skoro sve što nas zanima, ali to nije nimalo lak posao. Živimo u vremenu u kojem smo svakodnevno bombardovani velikom količinom informacija koje je nemoguće kvalitetno prihvatiti, pregledati, obraditi i iskoristiti na pravi način.

Treće poglavlje započinje iznošenjem osnovnih osobina internet pretraživača. Dat je kratak pregled kako rade internet pretraživači, njihov razvoj, uticaj Web mastera na rad internet pretraživača, kao i problemi internet pretraživanja.

Ubrzanim rastom Web-a korisnici su često suočeni sa problemom preopterećenja informacijama. Taj problem se može okarakterisati na dva načina:
- filtriranje informacija (Information filtering – svakodnevno smo suočeni sa velikim brojem informacija (e-mail, novosti na Web-u,...) od kojih su samo pojedine relevantne i važne, problem se javlja zbog lošeg upravljanja dobijenim informacijama),
- sakupljanje informacija (Information gathering – informacije se neprikladno pohranjuju što otežava njihovo kasnije pronalaženje zbog prevelikog broja drugih informacija).

Web sadržaji brzo se razvijaju i menjaju, ali i postaju zastareli. Kako količina i dostupnost raznolikih informacija svakodnevno raste, tako raste i potreba za što boljim snalaženjem s njima. Jedno rešenje ovog problema je primena inteligentnih agenata za pretraživanje Web-a. Devedesetih godina prošlog veka intezivno se radilo na području veštačke inteligencije i to u vezi tzv. inteligentnih agenata za pretraživanje Web-a.

Četvrto poglavlje započinje iznošenjem osnovnih osobina agenata. Tako je dat kratak pregled definicija agenata iz kojih proizilaze i različiti pristupi ovom području, osnovna svojstva inteligentnih agenata i elementarna arhitektura agenta. Potom se iznosi osnovni koncept rada pretraživačkih sistema kod kojih prikupljanje podataka obavljaju agenti za pretraživanje Web-a.

Nakon toga se u petom poglavlju upoznajemo sa pojmom inteligentnog agenta za pretraživanje Web-a. U literaturi se ujedno koriste i nazivi Web pauci (eng. Web spider), Web puzači (eng. Web crawler) i Web roboti. Na kraju poglavlja dat je opis izgradnje jednostavnog inteligentnog agenta za pretraživanje Web-a.

Šesto poglavlje opisuje metode koje agenti koriste za pretraživanje Web-a. Data je osnovna podela metoda na slepe i usmerene kao i njihove karakteristike. Navode se pretraživanja koja spadaju u slepe metode i dat je njihov opis. Potom se navode pretraživanja koja spadaju u usmerene metode i za svaku pojedinu metodu pretraživanja dat je opis same tehnike pretraživanja.

Sedmo poglavlje predstavlja case study Google-a. Date su osnovne karakteristike Google internet pretrazivača, njegova arhitektura, veličina indeksa, način upotrebe, kao i neke dodatne činjenice o Google-u. Opisan je i Google Chrome, kao i Google-ov inteligentni agent GoogleBot.

2. Metodologija Naučnog Istraživanja

2.1. Predmet istraživanja

Predmet našeg istraživanja predstavljaju internet pretraživači, inteligentni agenti, njihov način funkcionisanja, osobine, struktura, kao i mogucnosti primene inteligentnih agenata kod internet pretraživača.

2.2. Hipotetički okvir istraživanja

Hipotetički okvir istraživanja sastoji se iz sledećih hipoteza:

Generalna hipoteza:

Razvojem interneta i internet pretraživača, došlo je do potrebe korišćenja inteligentnih agenata koji će pomoći korisnicima da se snađu u moru informacija.

Posebna hipoteza:

- Dizajn Inteligentnog agenta je specifikacija akcija koju treba da izvrši za svaku moguću sekvencu opažaja.
- Glavne osobine inteligentnih agenata i internet pretraživača
- Svaki inteligentni agent se sastoji iz dva dela i to arhitekture i programa
- Kako agent radi zavisi od programa i algoritama

2.3. Ciljevi i zadaci istraživanja

Naučni cilj ovog rada je predstavljanje internet pretraživača, inteligentnih agenata, otkrivanje sposobnosti internet pretraživača za pomoć ekspertskim sistemima u predviđanju korisničkih zahteva, odnosno efikasnom pretraživanju informacija i selekciji relevantnih sadržaja.

Društveni cilj ovog rada je predstavljanje korišćenje internet pretraživača (web search engine) kao mehanizma za efikasno pretraživanje (inteligentnih agenata), selekciju relevantnih sadržaja i dobijanje brzih odgovora u cilju konstantnog razvoja proizvoda, dobara i usluga.

2.4. Metode i tok istraživanja rada

U ovom radu, da bi se postigli postavljeni ciljevi kao i što efikasniji rezultat koji se dobija povezivanjem i odgovarajućom kombinacijom više naučnih metoda, korištene su sledeće metode:

- Metoda deskripcije
- Komparativna metoda
- Metoda analize
- Metoda sinteze

Metoda deskripcije, definisana kao postupak jednostavnog opisivanja ili ocrtavanja činjenica, procesa ili predmeta, te njihovih empirijskih potvrdjivanja odnosa i veza, ali bez naučnog tumačenja i objašnjavanja, je korišćena kroz ceo rad.

Komparativna metoda, koja podrazumeva upoređivanje istih ili srodnih činjenica, pojava, procesa i odnosa, odnosno utvrđivanja njihove sličnosti u ponašanju i intenzitetu kao i razlika među njima.

Metoda analize, je postupak naučnog istraživanja i objašnjavanja stvarnosti putem raščlanjivanja složenih misaonih tvorevina (pojmova, sudova i zaključaka) na njihove jednostavnije delove i elemente, kao i izračunavanje svakog dela (elementa) za sebe i u odnosu na druge delove celine.

Metoda sinteze, je suprotna metoda od analize koja se takođe primenjivala u izradi ovog rada.

3. Internet pretraživači – metode i analiza

3.1. Kako rade internet pretraživači

Kada se internet pretražuje preko nekog pretraživača, vrši se pretraga 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:

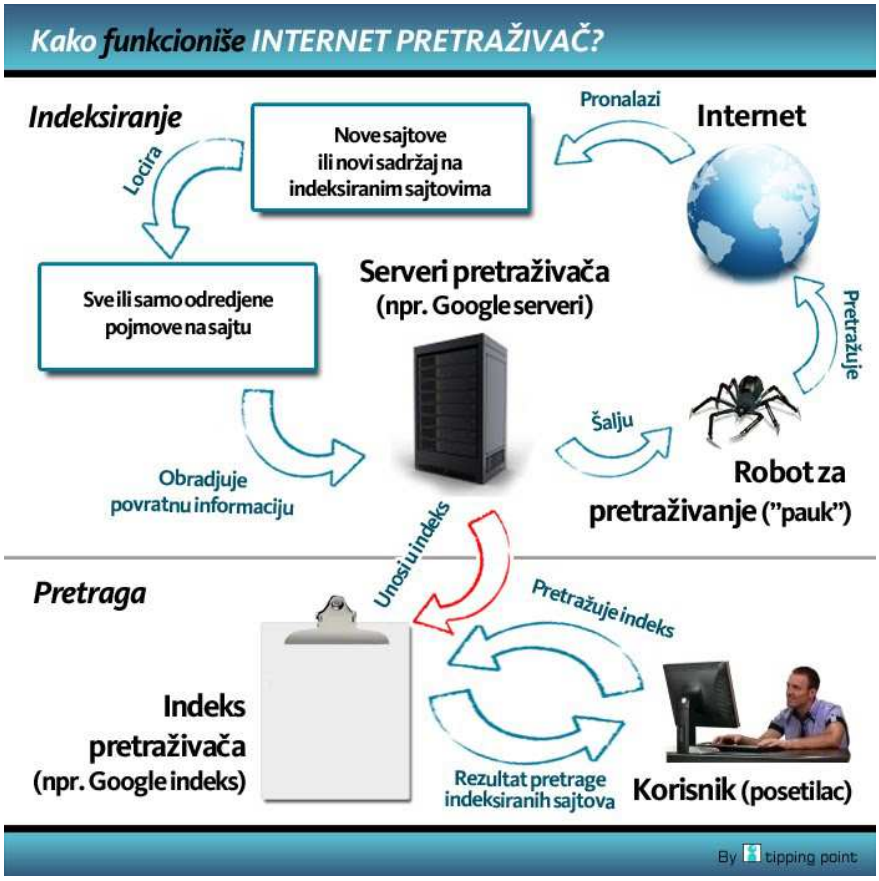
Crawlers - (Spiders, Robots) su pretraživači koji automatski "krstare" web-om, indeksiraju i ažururaju svoje baze podataka ("listnigs") pa korisnici kasnije pretražuju po njima. Znači, sve je bazirano na algoritmu indeksiranja - što je algoritam kvalitetniji, baza je optimizovanija. (Google, AltaVista)

Human Directrories - 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 koji se može downloadovati i startovati kao aplikacija, a ne kroz web browser, kao npr Copernic. (www.copernic.com)[1]

[1] http://en.wikipedia.org/wiki/Search_engine_technology



Slika 1. Kako rade internet pretraživači

Pretraživanje danas predstavlja ključni način pristupa sadržaju na Web-u. Usled enormnih dimenzija (preko 8 milijardi stranica), pristup informacijama na Web-u, direktnim posećivanjem svake od stranica korišćenjem njenog URL-a, ograničava korisnika isključivo na skup unapred poznatih stranica, što predstavlja gotovo zanemarljiv segment celokupnog Web-a. Jedan od načina koji omogućava efikasniji pristup, predstavljaju tzv. direktorijumi (Yahoo!, dmoz..), koji podrazumevaju kataloge web stranica, klasifikovane po tematici. Iako inicijalno popularan, ovakav pristup se ubrzo pokazao kao neefikasan, prvenstveno usled ogromne brzine ekspanzije Web-a, koja uzrokuje novim sadržajima, čije je pronalaženje i klasifikacija (koja se najčešće obavlja manuelno), izuzetno zahtevan posao. Dodatno, sa sve većim povećavanjem količine informacija, klasifikacija se pokazuje kao nedovoljno efikasan pristup za rešavanje problema preopterećenosti informacijama (information overload).

Najefikasniji vid pristupa sadržaja na Web-u, jeste korišćenje pretraživanja. Na ovaj način, izbegava se potreba za sistematskom organizacijom celokupnog Web-a, već se deo problema prenosi na samog korisnika, koji svojim upitima, vrši selekciju segmenata kojima želi da pristupi. Zadatak Web pretraživača (web search engine) je da pruži mehanizme koji korisniku omogućavaju efikasno pretraživanje i pomoć pri selekciji relevantnog sadržaja. U poslednjoj deceniji, razvijen je značajan broj izuzetno efikasnih pretraživača, od kojih neki danas predstavljaju najposećenjije lokacije na Web-u (Google, Yahoo!,Search,Ask..).

3.2. Internacionalno tržište

Udeo internet pretraživača na tržištu se razlikuje od tržišta do tržišta, kao i konkurencija. 2003. godine, Danny Sullivan je izjavio da Google zastupa približno 75% svih internet pretraga. Na tržištima izvan SAD, Google udeo u internet pretraživačima je često i veći, i Google je i danas dominantni internet pretraživač. 2006. godine, Google je pokrio oko 40% tržišta u SAD, ali je zato u Nemačkoj bio prisutan sa 85%-90% udela na tržištu. U to vreme je bilo više stotina SEO firmi u SAD, a samo pet u Nemačkoj.

U Rusiji je situacija obrnuta. Lokalni internet pretraživač "Yandex" kontroliše 50% tržišta, dok Google ima manje od 9% udela. U Kini "Baidu" internet pretraživač i dalje vodi u tržišnom udelu, ali od 2007. godine mu se Google sve više približava.

Uspešna optimizacija internet pretraga za internacionalno tržište zahteva profesionalan prevod web stranica, registraciju na domenu koji je trenutno vodeći na ciljnom tržištu, kao i web hosting sa lokalnom IP adresom. Osim toga, osnovni elementi optimizacije su u osnovi isti, bez obzira na jezik.[2]

[2] Tara Calishain (2007.): „Web Search Garage“

3.3. Razvoj internet pretraživača

Web masteri i provajderi sadržaja su počeli optimizaciju sajtova za internet pretraživače sredinom 90-ih godina, kada su prvi internet pretraživači katalogizirali Web. U početku, sve što je web-master trebalo da uradi je da pošalje stranicu, ili URL, raznim internet pretraživačima koji bi potom poslali "spider" (pauka) da pretražuje po toj stranici, ekstraktuje linkove ka drugim stranicama sa te strane, i vrati informacije pronadjene na stranici za indeksiranje. Sam proces uključuje download stranice uz pomoć pauka i pohranjivanje iste na serveru internet pretraživača, gde drugi program, poznat kao "indexer", ekstraktuje razne informacije o stranici, kao sto su reči koje sadrži i gde su te reči locirane, takodje značaj specifičnih reči, i naravno sve linkove koje stranica sadrži, koji se zapisuju u raspored za kasniji "crawling".

Ranije faze algoritama pretrage su se oslanjale na informacije dobijene od web-mastera kao sto su ključne reči, meta tagovi ili indeksirani fajlovi kod internet pretraživača kao sto je "ALIWEB". Meta tagovi služe kao vodič sadržaja neke strane. Ali ispostavilo se da korišćenje meta tagova za indeksiranje stranica nije dovoljno pouzdano jer izbor ključih reči, od strane web mastera, u meta tagovima može potencijalno da bude neprecizan prikaz stvarnog sadržaja sajta. Neprecizni, nepotpuni, i konzistentni podaci u meta tagovima su dovodili do rankiranja stranica u nerelevantnim pretragama. Provajderi web sadržaja su takodje manipulisali raznim atributima unutar HTML koda stranica sa namerom da se dobro rankiraju u internet pretragama.

Oslanjajući se u velikoj meri na faktore isključivo kontrolisane od strane web mastera, rani internet pretraživači su imali problema tipa iskorišćavanja i manipulacije rankiranja. Radi pružanja boljih rezultata svojim korisnicima, internet pretraživači su morali da se adaptiraju da bi osigurali da njihove pretrage prikazuju najviše relevantnih rezultata, umesto nerelevantnih stranica natrpanih brojnim ključnim rečima od strane beskrupuloznih web-mastera. Obzirom da se uspeh i popularnost internet pretraživača odredjuje po mogućnostima istog da proizvede najviše relevantnih rezultata za bilo koju pretragu, dozvoljavanje da ti rezultati budu netačni bi navelo korisnike da potraže druge izvore pretraživanja. Internet pretraživači su na to odgovorili razvijanjem kompleksnijih algoritama za rankiranje, uzimajući u obzir dodatne faktore kojima su web-masteri manipulisali.

Dok su bili studenti poslednje godine na Stanford Univerzitetu, Larry Page i Sergey Brin su razvili "backrub", internet pretaživač koji se oslanjao na matematički algoritam da oceni prominentnost web stranica. Broj koji je algoritam izračunavao, "PageRank", je funkcija kvantiteta i snage dolaznih linkova. "PageRank" ocenjuje koja je verovatnoća da će web korisnik koji nasumice surfuje doci do posmatrane stranice, i prati linkove sa strane na stranu. Praktično, ovo znači da su neki linkovi jači od drugih, odnosno stranica bolje rankirana "PageRank" algoritmom ima više šanse da se do nje dodje nasumičnim surfovanjem.[3]

[3] http://en.wikipedia.org/wiki/Search_engine

Page i Brin su osnovali Google 1998.godine. google je privukao lojalnu pratnju iz sve većeg broja internet korisnika, kome se svideo njegov jednostavan dizajn. "Off-page" faktori (kao sto su "PageRank" i hiperlink analiza) su kao i "on-page" faktori (kao sto su učestalost ključnih reči, meta tagovi, hederi, linkovi i struktura sajta) smatrani zaslužni što Google uspešno izbegava manipulaciju vidjenu kod internet pretraživača koji uzimaju u obzir samo "on-page" faktore za ranking stranica. Iako je bilo teže izmanipulisati "PageRank', web-masteri su razvili alat za pravljenje linkova i šema da bi manipulisali "Inktomi" internet pretraživačem, i ove metode su se pokazale delimično primenjive na manipulaciju "PageRank". Mnogi sajtovi su se fokusirali na razmenu, kupovinu i prodaju linkova, često u ogromnim količinama. Neke od ovih šema, ili "link farms" su uključivale kreiranje nekoliko hiljada sajtova radi jedinstvene svrhe spamovanja linkova. Poslednjih godina najveći internet pretraživači su počeli više da se oslanjaju na "off-web" faktore kao sto su godište, pol, lokacija, i istorija pretrage ljudi koji vrše iste, da bi dodatno poboljšali rezultate.

Do 2007. godine, internet pretraživači su integrisali širok spektar neobjavljenih faktora u njihove algoritme rankiranja da bi se smanjio uticaj manipulacije linkova. Google koristi vise od 200 različitih signala. Tri vodeća internet pretraživača Google, Yahoo, Micrisoft's Live Search, ne otkrivaju algoritme koje koriste za rankiranje stranica. Poznati SEO strucnjaci, kao sto su Rand Fishkin, Barry Schwartz, Aaron Wall i Jill Whalen, su proučili različite pristupe optimizaciji internet pretrage, i izdali su svoja zapažanja na online forumima i blogovima, SEO stručnjaci takodje proučavaju patente na kojima se razni internet pretraživači zasnivaju da bi stekli uvid u nijhove algoritme.

3.4. Optimizacija internet pretrage

Optimizacija intenet pretrage (Search engine optimization ili SEO) je proces poboljšanja količine i kvaliteta saobraćaja od internet pretraživača ka web sajtu putem "prirodnih" ("organskih" i "algoritamskih") rezultata pretrage. Tipično, što se sajt ranije pojavi u listi pretrage, više posetilaca će isti posetiti putem linka internet pretraživača. SEO se može koristiti u različitim vrstama pretraga, uključujući pretragu slika, loklanu pretragu i specifičnu za industriju, vertikalnu pretragu.

Kao strategija internet matketinga, SEO uzima u obzir način na koji internet pretraživači rade i šta korisnici traže. Optimizacija web sajta primarno uključuje editovanje sadržaja istog i HTML kodiranje da bi se poboljšala relevantnost sa specifičnim ključnim rečima i da bi se uklonile barijere aktivnostima indeksiranja internet pretraživača.

Akronim "SEO" takodje može da se odnosi na "search engine optimizers" (ljudi koji se bave poslom optimizacije internet pretrage), pojam koji je usvojen od strane industrije konslultanata koji vrše optimizaciju projekta na zahtev klijenata, i od strane zaposlenih koji pružaju SEO usluge unutar firmi. Ti ljudi mogu SEO uslugu da ponude samo za sebe ili kao deo šire markentinške kampanje. Zbog postojanja mogućnosti da efektivan SEO može da zahteva promene u HTML kodu sajta, SEO taktika može biti inkorporisana u razvoj i dizajn web sajta. Pojam "search engine friendly" (u dobroj saradnji sa internet pretraživačima) se može koristiti da opiše dizajn web sajta, menije, sisteme upravljanja sadržajem i kupovine korpe je lako optimizovati.[4]

Jos jedna klasa tehnika, poznata kao "black hat SEO" ili "spamdexing", koristi metode kao sto su "link farms" i "keyword stuffing" koje degradiraju relevantnost rezultata pretrage kao i korisnikovo iskustvo sa internet pretraživačima. Internet ptreraživači traže sajtove koji koriste ovu tehniku sa ciljem uklapanja istih iz indeksa.

[4] http://sr.wikipedia.org/sr/SEO_optimizacija_veb_sajta

3.5. Uticaj Web-mastera na rad internet pretraživača

Do 1997. godine internet pretraživači su uvideli da web-masteri rade na što boljem rankiranju njihovih sajtova u pretragama, kao i da neki web-masteri manipulišu svojim rankom u rezultatima pretraga time što natrpavaju stranice suvišnim ili nerelevantnim ključnim rečima. Rani internet pretraživači, kao sto je "Infoseek", su prilagodili svoje algoritme u pokušaju da spreče web-mastere da manipulišu rankiranjem.

Zbog visoke tržisne vrednosti odredjenih rezultata pretrage, prisutan je potencijal za neprijateljske odnose izmedju internet pretraživača i optimizera. 2005. godine je kreirana godišnja konferencija, "AIRWeb", "Adversarial Information Retrieval" na internetu, u cilju diskutovanja i minimiziranja štetnih efekata agresivnih provajdera web sadržaja.

SEO kompanije koje koriste previše agresivne tehnike rizikuju da sajtovi njihovih klijenata budu isključeni iz rezultata pretrage. Na primer, 2005. godine, "Wall Street Journal" je izveštavao o kompaniji, "Traffic Power", koja je navodno koristila viskoko rizične tehnike, a nije izložila taj rizik svojim klijentima. "Wired" magazin je objavio da je ista kompanija tužila blogera i SEO srtučnjaka "Aaron Wall" zbog pisanja o isključenju iz pretraga. Matt Cutts iz Google je kasnije potvrdio da je Google isključio "Traffic Power" i neke od njihovih klijenata iz pretraga.

Neki internet pretraživači su se umešali u SEO komapnije, i česti su sponzori i gosti SEO konferencija, četova, i seminara. U stvari, sa pojavom plaćenog učešća, neki internet pretraživači sada imaju interese za opšte dobro optimizacijske zajednice. Najveći internet pretraživači pružaju informacije i vodiče da bi pomogli pri optimizaciji sajtova.

Google ima "Sitemaps" program koji omogućuje web-masterima da vide da li Google ima nekih problema pri indeksiranju njihovih sajtova i takodje pruža podatke o Google saobraćaju ka tim sajtovima. "Google guidelines" je lista predloženih tehnika koje je Google predložio kao vodič web-masterima. "Yahoo! Site Explorer" pruža način web-masterima da pošalju URL, da utvrde koliko stranica se nalazi u Yahoo! Indeksu i da pregledaju informacije o linkovima.[5]

[5] http://en.wikipedia.org/wiki/Search_engine_optimization

3.6. Dodatne činjenice o internet pretraživačima

Pre nego što su postojali internet pretraživači bila je kompletna lista svih web servera. Kako se pojavljivalo sve više web servera centralna linija nije mogla više sve to da prati. NCSA web sajt je objavio nove web sajtove pod naslovom "What's New!" ali više nije postojao kompletan listing.

Prva alatka za pretraživanje pre web-interneta bila je "Archie". Ime je dobijeno tako što se od reci "archive" izbacilo slovo "v". "Archie" je 1990. godine stvorio Alan Emtage. Ovaj program je skidao listinge direktorijuma svih fajlova lociranih na javnim anonimnim FTP sajtovima, stvarajući pretražljivu bazu podataka imena fajlova; ali ipak Archie nije indeksirao sadržaj ovih sajtova.

Uspon "Gopher" 1991.godina, vodio je ka dva nova programa za pretragu, "Veronica" i "Jughead". Kao Archie, ovi programi su tražili imena fajlova i naslove pohranjene u Gopher indeksnim sistemima. Veronica (**V**ery **E**asy **R**odent-**O**riented **N**et-wide **I**ndex to **C**omputerized **A**rchives) je pružala pretrage ključnih reči većine Gopher meni naslova iz celokupnog Gopher listinga. Jughead (**J**onzy's **U**niversal **G**opher **H**ierarchy **E**xcavation **A**nd **D**isplay) je bio alatka za pretragu informacija menija sa odredjenih Gopher servera.

U junu 1993.godine, Matthew Gray, napravio je verovatno prvog web robota, Perl baziran WWW Wanderer, i koristio ga je za generisanje indeksa zvanog "Wandex". Wanderer je imao svrhu da izmeri veličinu WWW, što je i uradio do kraja 1995.godine. Internet pretraživač "Aliweb" se pojavio 1993.godine. On nije koristio web robota, već je zavisio od web administratora da ga obaveste o postojanju i formatu indeks fajla za svaki web sajt.

"JumpStation" 1993. godina, je koristio web robota da pronadje web stranice i da izradi svoj indeks, i koristio je web formu kao interfejs za svoj program za upite. Samim tim je bio prvi WWW alat za pretragu koji je kombinovao tri najvažnije karakteristike internet pretraživača ("crawling", "indexing", i "seraching").

Jedan od prvih "full text" "crawler"-baziranih internet pretraživača je bio "WebCrawler", koji se pojavio 1994. godine. Za razliku od svojih prethodnika, on je dozvoljavao korisnicima da traže bilo koju reč na bilo kojoj web strani, što je postalo standard za sve velike internet pretraživače. Takodje je bio prvi internet pretraživač poznat širokoj javnosti. Web crawler je kompijuterski program koji pretražuje na metodičan i automatizovan način. Drugi nazivi za web crawlere su: ants (mravi), automatic indexers (automatizovani indekseri), bots (botovi), worms (crvi), web spiders (web pauci), i web roboti. Sam proces u kom oni učestvuju se zove "web crawling" ili "spidering". Mnogi web sajtovi, naročito internet pretraživači, koriste spidering kao sredstvo za pružanje svežih podataka. Web crawleri se uglavnom koriste za kreiranje kopija svih posećenih stranica radi kasnijeg procesiranja od strane internet pretraživača, koji će indeksirati preuzete stranice da bi pružili brže pretrage. Crawleri se takodje mogu koristiti za automatizovanje održavanja web sajta, kao što je provera linkova ili validnost html koda. Pored toga crawleri se mogu koristiti i za prikupljanje informacija odredjenih tipova sa web strana, kao što su e-mail adrese.

Ubrzo potom su se pojavili mnogi internet pretraživači i borili se za popularnost. Medju njima su: "Magellan", "Excite", "Infoseek", Inktomi", "Northern Light" i "Alta Vista". "Yahoo" je bio medju najpopularnijim načinima za ljude da pronadju web stranice od interesovanja, ali njegova pretraga je bila zasnovana na sopstvenom web direktorijumu, umesto na "full-text" kopijama web stranica. Korisnici su bili u mogućnosti da pregledaju direktorijum umesto da vrše pretragu baziranu na nekoj ključnoj reči.

1996. godine, "Netscape" je hteo da sklopi ugovor sa nekim internet pretraživačem da bude njihov ekskluzivni internet pretraživač. Interesovanje je bilo toliko veliko da je postignut dogovor sa 5 velikih internet pretraživača, gde su za 2 miliona dolara godišnje naknade svaki od njih menjali u rotaciji na Netscape sajtu. Tih 5 internet pretraživača su bili: Yahoo, Magellan, Lycos (1994. godine je postao ogroman komercijalni poduhvat), Infoseek, i Excite.

Internet pretraživači su bili takodje poznati kao najsjajnije zvezde interneta i najbolja prilika za ulaganje za vreme velikog boom-a internet tehnologija koji se desio krajem prošlog veka. Nekoliko kompanija su spektakularno ušle na tržište, ostvarujući rekordne dobitke pri prvom javnom izdavanju akcija. Neke kompanije su ugasile svoje javne internet pretraživače, i sada postoje samo kao marketinška komercijalna izdanja, kao sto je "Northern Light".

2000. godine se " Google" internet pretraživač istakao. Postizao je bolje rezultate u mnogim pretragama uz pomoć inovacije nazvane "PageRank". Ovaj interativni algoritam rankira web stranice u zavisnosti od broja i ranka drugih web sajtova i stranica koje poseduju link ka posmatranoj stranici, zasnovano na premisi da se linkovi ka dobrim ili poželjnim stranicama postavljaju češće od ostalih. "Google" je takodje zadržao minimalistički izgled interfejsa, drugačije od većine konkurencije koji su integrisali svoje internet pretraživače u web portal.

Do 2000. godine "Yahoo!" je pružao usluge bazirane na "Inktomi" internet pretraživaču. "Yahoo" je kupio "Inktomi" 2002. godine. "Yahoo" je prešao na "Google" sistem pretraživanja do 2004. godine, kada je pustio u rad sopstvenu tehnologiju internet pretraživanja baziranu na kombinovanim tehnologijama pretraživača.

Microsoft je prvo lansirao "MSN Search" 1998. godine koristeći rezultate pretraga od "Inktomi". "Inktomi" je početkom 1999. godine počeo da prikazuje listinge sa "Looksmart" spojene sa rezultatima sa "Inktomi" osim u malom vremenskom periodu u toku 1999. godine kada su korišćeni rezultati pretraga od "AltaVista". 2004. godine Microsoft je počeo transfer na sopstvenu tehnologiju pretraživanja, koja je koristila njihov web "crawler".

Od druge polovine 2007. godine "Google" je ubedljivo najpopularniji internet pretraživač širom sveta. Takodje su i odredjeni broj internet pretraživača specificnih za odredjenje zemlje postali poznati; na primer tu su "Baidu" koji je veoma popularan u Kini i "guruji.com" iz Indije.

Internet pretraživači se mogu podeliti na dve grupe: meta pretraživači i "pravi" Internet pretraživači.

Meta-pretraživači su vrsta internetskog pretraživača uz pomoć čijeg učešća je moguće pretraživati nekoliko drugih izvora informacija (najčešće pretraživača) odjednom. Korisnici upit unose samo jednom, a on se potom distribuira prema više drugih pretraživača i/ili baza podataka, uz zajednički prikaz rezultata sa svakog od njih. Meta pretraživači su programi koji skupljaju informacije o web stranicama koje se nalaze u META oznakama HTML dokumenata.

Pravi Internet pretraživači su programi koji skupljaju sve dostupne HTML dokumente web stranica i sortiraju i obrađuju iste u svojim bazama podataka. Na web stranicama ovih pretraživača možemo pogledati skupljene informacije u bazi podataka.[6]

[6] John Battelle (2005.): „The Search“

3.7. Problemi internet pretraživanja

Iako se svakodnevno unapredjuju, internet pretraživači su u osnovi na niskoj grani tehnološkog razvoja. Neki čitaju meta tagove , neki samo prvih par stotina reči iz teksta. Neki isčitavaju i jedno i drugo i povezuju ih zajedno, što u prevodu znači da ako web sajt ima meta tagove koji nisu u direktnoj vezi sa sadržajem istog, taj web sajt neće biti visoko rankiran.

Ako web stranica počinje sa velikim slovom, postoji mogucnost da će internet pretraživač iščitati samo to i da ce ignorisati stvarni sadržaj koji dolazi mnogo karaktera posle.

Takodje je bitno uzeti u obzir da internet pretraživači iščitavaju stranice onako kako one izgledaju u HTML-u, a ne kako se one prikazuju u browser-u. To može da predstavlja potencijalni problem u slučaju korišćenja tabela, jer će internet pretraživač iščitati prvo alocirane karaktere, pre nego što stigne do stvarnog sadržaja.[7]

[7] http://www.cins.org.rs/?p=367

4. Inteligentni agenti, opšti pojmovi, arhitektura

World Wide Web je postao nepregledan izvor informacija. Pronalaženje tačno određene informacije je često vrlo teško zbog kompleksnosti organizacije i količine pohranjenih informacija. Informacije su pohranjene u dokumentima koji imaju svoja imena (web stranice) i nalaze se u određenim direktorijumima. Pronalaženje određene informacije zahteva poznavanje adrese na kojoj je smeštena, a pogoditi gde se nalaze informacije koje nam trebaju je gotovo nemoguće. Srećom, pretraživački sistemi poput Google-a, Alta Viste, Bing-a i Yahoo-a, i mnogih drugih, nam olakšavaju pretraživanje Web-a. Mnogi intenrnet pretraživači pretražuju web stranice, pa iz njih crpe podatke i dokumente kako bi zadovoljili korisnikove potrebe. Međutim, često rezultati tih pretraživanja nisu relevantni korisnikovom zahtevu. Upravo da bi se izbegao ovakav problem razvijeni su inteligenti agenti za pretraživanje Web-a.

Inteligentni agenti imaju takve osobine i mogućnosti da se smatraju novim pomakom u informatici. Koriste se širom sveta, kako u akademskim zajednicama tako i u poslovnim krugovima kod donošenja bitnih poslovnih odluka.

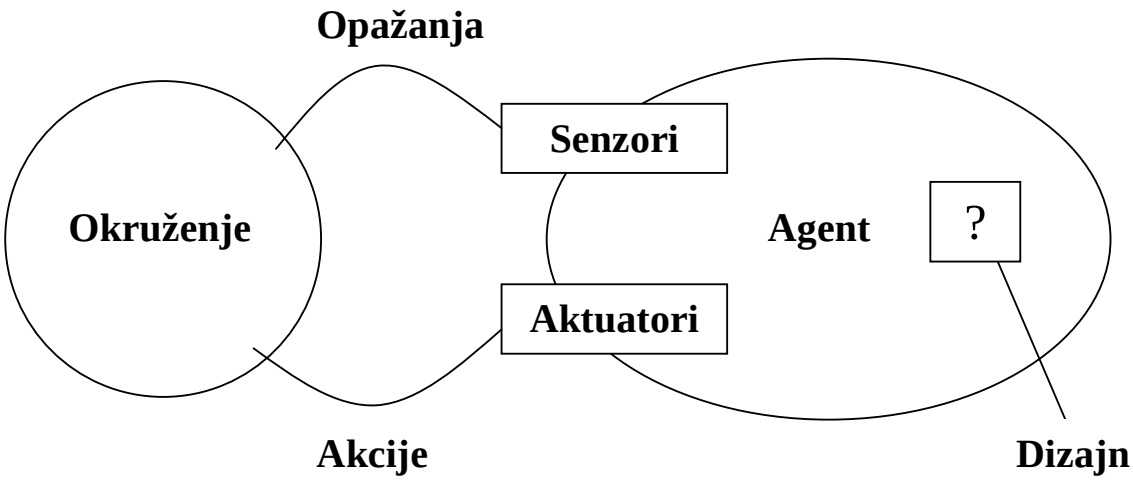
Područje inteligentnih agenata je vrlo mlado. Istraživanja ovog područja mogu se podeliti na razvijanje teorije o agentima, na razvijanje arhitekture agenata i na razvijanje jezika orijentisanih na agente, kako onih u kojima su agenti napisani tako i onih koji služe za komunikaciju među agentima. Uprkos preprekama koje postoje, inteligentni agenti predstavljaju izazov jer prednosti koje nude premašuju poteškoće koje stoje na putu njihove praktične primene.[8]

[8] http://members.tripod.com/~veliborb

4.1. Agenti

Šta je to agent? Uopšteno govoreći, agent je entitet koji nešto radi i to radi u ime nekoga. Definisati pojam agenta je podjednako teško kao i definisati pojam npr. objekat. Pojam je jednostavno preopširan. Postoje i mnogo određenije definicije šta su agenti. Međutim, sve definicije odražavaju raznolikost shvatanja pojma, što je posledica mladosti čitavog područja.

Intuitivno se pri spominjanju agenta podrazumeva objekat koji samostalno deluje u nekoj okolini, prilagođava se okolini u kojoj deluje, ima sposobnost percipiranja stanja u kojem se njegova okolina nalazi, delovanjem menja stanja okoline i ima sposobnost učenja.



Slika 1: Opšti izgled inteligentnog agenta[9]

[9] www.scribd.com/doc/39403486/inteligentni-agenti

4.2. Definicije agenata

Nekoliko definicija agenata:

- Pojam agent prikazuje dva ortogonalna koncepta:
1) sposobnost agenta da deluje autonomno tj. nezavisno od korisnika
2) sposobnost agenta da zaključuje o nekom području interesa.

- Agent je nešto što može prihvatati vlastitu okolinu kroz senzore i delovati na okolinu kroz efektore.

- Autonomni agenti su sistemi koji prihvataju kompleksnu dinamičnu okolinu, osećaju je i deluju autonomno u toj okolini te na taj način realizuju skup ciljeva zbog kojih su oblikovani.[10]

- Agenti su stalni programski entiteti namenjeni za posebnu svrhu. To znači da se agenti razlikuju od potprograma; agenti imaju svoje vlastite ideje o izvršavanju zadatka, svoj vlastiti raspored. Posebna svrha razlikuje agente od potpuno višefunkcionalnih aplikacija; agenti su manji.

[10] http://bs.scribd.com/doc/59442859/01-Vestacka-Inteligencija

4.3. Podela agenata

Rad agenta razmatra se kroz njegovo delovanje unutar njemu dodeljene okoline. Interakcijom agenta s okolinom menjaće se stanja okoline. Opšte je prihvaćena definicija agenata prema kojoj se agenti mogu podeliti na slabu i jaku podelu.[11]

Slaba podela agenata ima sledeća obeležja:

- Samostalnost (autonomous): Agent ima sposobnost da izvršava operacije samostalno bez potrebe za intervencijom korisnika. Ovo svojstvo zahteva od agenta posedovanje mehanizama koji će mu omogućiti samostalno dobijanje svih potrebnih informacija o okruženju u kojem se nalazi. Zavisno od dobijenih informacija agent izvršava pripadajuće operacije. U slučaju greške mehanizmi moraju omogućiti nastavak rada agenta.

- Društvenost (social ability): Agent mora posedovati mehanizme koji mu omogućuju koordinisanje svojih operacija s drugim agentima u svom okruženju. Obeležje društvenosti agenta ostvareno je korišćenjem komunikacijskih protokola, interakcijskih protokola i ontologija.

- Reaktivnost (reactivity): Reaktivni agenti ne poseduju interni model koji im omogućava predviđanje budućeg stanja okoline, već reaguju isključivo na temelju sadašnjeg znanja i skupa akcija koje poznaju. Oni reaguju na uticaje iz okoline s akcijom predviđenom za trenutno stanje okoline i vrstu uticaja.

- Proaktivnost (proactiveness): Za razliku od reaktivnih agenata, koji samo reagiraju na uticaje iz okoline, proaktivni agenti poseduju mehanizme koji im omogućavaju uticanje na okolinu menjajući pritom njeno stanje. Pri tome iskazuju ciljno orijentisano ponašanje, odnosno, svaki agent ima jedan ili nekoliko ciljeva. U situacijama kada stanje okoline nije u skladu s njegovim ciljevima, agent utiče na okolinu dok ne ostvari zadati cilj.

U području veštačke inteligencije vrlo je često zastupljena jaka podela agenta koja ima sledeća obeležja:

- Pokretljivost (mobile): Agent se može kretati s jednog mrežnog čvora na drugi. Svaki pokretni agent sastoji se od tri komponente: (1) programskog koda koji sadrži logiku agenta, (2) podataka, odnosno, internih atributa koji predstavljaju znanje koje agent poseduje i (3) stanje izvođenja

- Racionalnost (rationality): definiše da ako agent ima skup ciljeva, od kojih je samo jedan aktivan, on neće izvoditi akcije koje bi mogle biti u suprotnosti s njegovim trenutnim ciljevima. Racionalan agent uvek mora izvoditi akcije koje bi maksimizirale očekivani rezultat pri tome koristeći svoje znanje o trenutnom i budućem stanju okoline

- Dobronamernost (benevolence): Agentovi ciljevi ne smeju biti međusobno konfliktni, ako se od agenta želi da maksimizira očekivani rezultat.

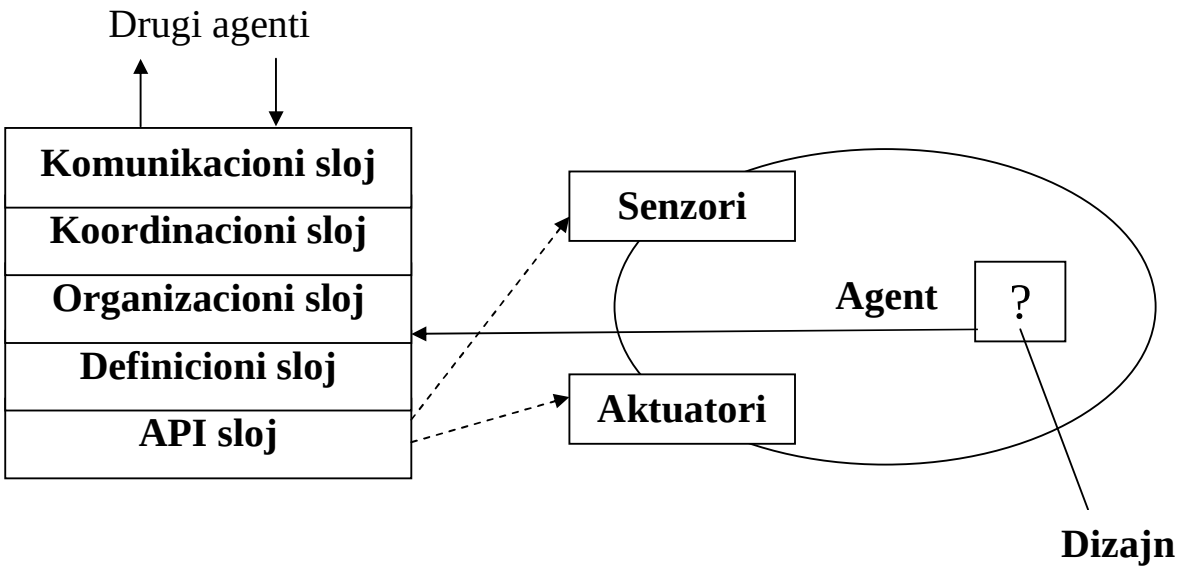
[11] www.csc.liv.ac.uk/~mjw/pubs/ker95/ker95-html.html

4.4. Inteligentni agenti

Često se umesto pojma agent koristi pojam inteligentni agent čime se naglašava da agent mora posedovati i određenu količinu inteligentnog ponašanja pri obavljanju svojih zadataka. Definicija inteligencije je sigurno složenija od definicije agenta. Inteligencija se opisuje različitim atributima poput sposobnosti učenja, razumevanja, planiranja i predviđanja, rešavanja problema, stvaranja novih ideja i modeliranja spoljnog sveta. Inteligenciju možemo opisati kao posedovanje mehanizama ponašanja usmerenih rešavanju zadanog cilja.[12]

Inteligentni agenti sadrže određeni stupanj inteligencije (nivo zaključivanja i učenja odnosno sposobnosti da prihvate korisnikove izjave o ciljevima i da obave zadatke koji se postave pred njih). Inteligentni agenti su agenti koji prate naše ponašanje i na temelju njega su u stanju naučiti kako mi rešavamo probleme i sl. Kad su to naučili oni mogu te zadatke obavljati samostalno, uz povremene provere vlasnika kod nejasnih ili novih situacija.

Svojstva inteligentnih agenata:
- učenje – od korisnika, od drugih agenata, iz ostalih izvora
- saradnja – rad s drugim agentima kako bi postigao svoj cilj
- mobilnost – pokretljivost agenata po mreži, izvođenje na različitim računarima
- personalizacija – poznavanje svog korisnika, njegovih interesa i preferencija
- adaptibilnost – uče iz različitih izvora i iz korisničkih akcija.



Slika 2: Opšti unutrašnji izgled inteligentnog agenta

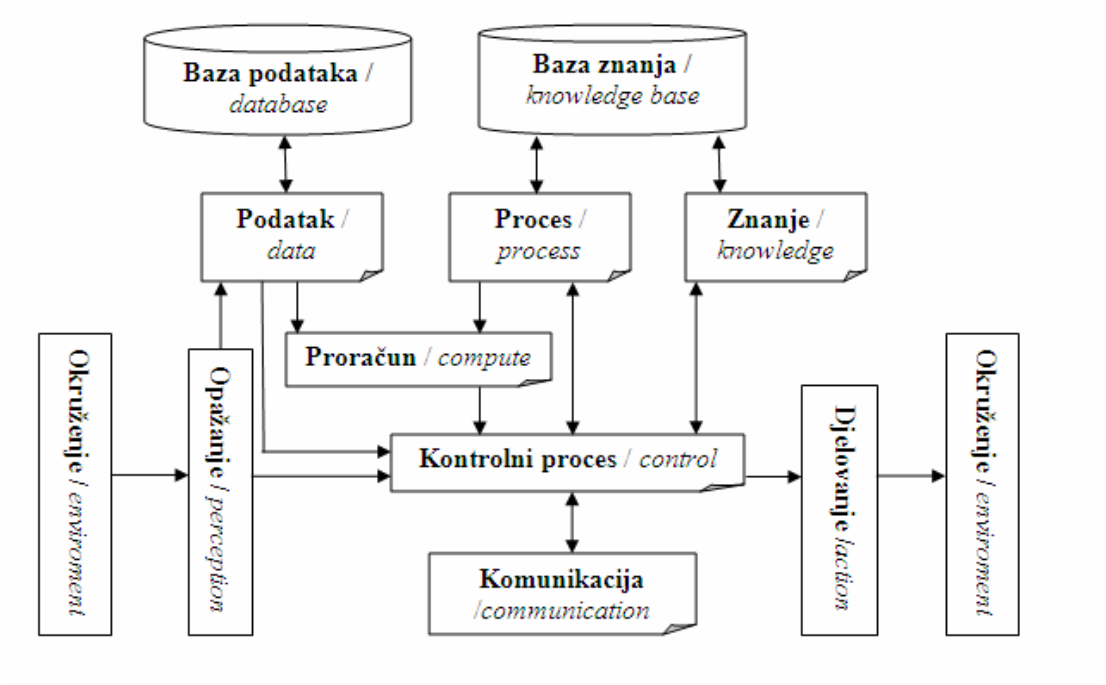
[12] http://internet.fon.bg.ac.rs/download/piai/Agents%20-%20Slides.pdf

4.5. Arhitektura agenata

Kad se govori o arhitekturi agenata misli se na takvu konstrukciju računarskog sistema koja omogućava agentima da to zaista i budu (da imaju osobine koje ih čine agentima). Arhitektura agenata je određena metodologija za izgradnju agenata, ona određuje kako se agent dekomponuje u niz modula i kako su ti moduli povezani. Ona pokazuje i način na koji moduli obavljaju svoje aktivnosti, a u svrhu obavljanja zadatka. U arhitekturi agenata mogu se i pokazati veze između različitih agenata, a ne samo struktura jednog agenta. Kod arhitekture ne može se govoriti o jednoj, jedinstvenoj arhitekturi agenta. Kakva će ona biti zavisi od svrhe agenta. Postoje određeni zahtevi koji se postavljaju pred arhitekturu agenata. Pre svega ona bi trebala omogućiti autonomiju, komunikaciju, učenje, orijentisanost ka cilju, mobilnost i stabilnost. Postoje različiti nivoi arhitekture u smislu da određeni agent ne mora imati sve osobine, ali one moraju biti iz gore navedenog skupa kako bi se entitet smatrao agentom. Jedan od najčešćih zahteva za arhitekturu jeste da omogući samostalnost. Učenje i zaključivanje drugi su važan aspekt koji mora osigurati arhitektura agenta. Budući da agent deluje u okolini koja sadrži i druge agente i različite usluge, potrebna je takva arhitektura koja standardizuje komunikaciju među agentima i omogućuje pristup određenim uslugama.[13]

Arhitektura agenta je zapravo struktura njegovog programa. Model samostalnog agenta (u smislu ostvarenja cilja) sastoji se od 8 jedinica i uključuje:
- jedinicu za opažanje (perception unit),
- jedinicu za obradu (process unit),
- kontrolnu jedinicu (control unit),
- radnu jedinicu (action unit),
- jedinicu za komunikaciju (communication unit),
- jedinicu znanja (knowledge unit),
- računsku jedinicu (compute unit),
- jedinicu sa podacima (data unit).

[13] Nguyen N.T., L.C. Jain, (2009): „Intelligent Agents in the Evolution of Web and Applications“



Slika 3: Arhitektura agenta

Jedinica za opažanje očitava podatke iz agentovog okruženja. Ona sadrži listu pojedinačnih stanja koja definišu ukupno stanje okruženja. Ukoliko se neko od pojedinih stanja promeni, lista podataka se ažurira, zatim se obaveštava kontrolna jedinica o promenama kako bi preduzela odgovarajuće mere. Jedinica za obradu sadrži cilj ili više ciljeva te njihove međusobne odnose. Radna jedinica sadrži sve delatnosti koje je agent sposoban obaviti. Kontrolna jedinica odlučuje koje će se delatnosti obaviti radi ostvarenja cilja, koji takođe ona bira. Ukoliko se jedinica za opažanje uopšte promeni u okruženju, tada kontrolna jedinica aktivira reakcijska delovanja na te promene. U računskoj jedinici su definisane funkcije kojima se biraju ciljevi i optimalna delovanja. Takođe sadrži algoritme i mehanizme odabiranja optimalnih delovanja. Kontrolna jedinica poziva funkcije iz računske jedinice kako bi se odredio sledeći cilj i prikladna delovanja. Jedinica za razumevanje nadgleda deo odgovoran za znanje agenta koje se primenjuje pri rešavanju stvarnih problema. Jedinica s podacima omogućava pristup svim mehanizmima u bazu podataka. Komunikacijska jedinica definiše načine komunikacije između agenata. Danas se inteligentni agenti koriste pri obavljanju mnogih zadataka kao što su pretraživanja sadržaja na Web-u, razvrstavanja elektronske pošte, odabir korisniku zanimljivih priloga u novinskim grupama, podrška računarskim mrežama, pretraživanje baza podataka i prilagodjavanje korisničkog interfejsa raznim individualnim potebama korisnika.

5. Agenti za pretraživanje Web-a

Porast popularnosti i broja dostupnih sadržaja, koji su u potpunosti neorganizovani, ruši osnovni koncept Weba - jednostavno pronalaženje informacija. Zbog toga se počevši od 1994. g. razvija niz alata, čiji je jedini zadatak pronalaženje željenih sadržaja na Web-u. Ti alati se razvijaju u dva osnovna smera, pa razlikujemo kataloge i pretraživače (search engines). Jedni i drugi sadrže bazu web stranica, indexa, URL-ova (Uniform Resource Locator - adresa koja jedinstveno ukazuje na neki sadržaj na Internetu), a osnovna razlika je u načinu prikupljanja tih podataka. Kod kataloga, sadržaj baze isključivo zavisi od ljudi, tj. ljudi pretražuju Web, za interesantne stranice napišu sažetak i sažetak se uz URL adresu stranice sprema u bazu. Katalozi se koriste još i danas, a jedan od najpopularnijih je Yahoo. Za razliku od kataloga, prikupljanje podataka kod pretraživača je u potpunosti automatizovano i obavljaju ga agenti za pretraživanje Web-a (Web Searching Agent) koji su još poznati pod nazivom pauci, puzači, crvi (spider, crawler, worm).

Agenti za pretraživanje Web-a „lutaju“ Web-om u potrazi za novim stranicama, i kada ih pronađu „dovlače“ ih na računar i pohranjuju u bazu. Mada razni nazivi poput pauka, lutalica, puzača, crva stvaraju privid da se agenti gibaju, to nije istina, oni su stacionirani na računaru i na taj računar „dovlače“ stranice. Oni „skaču“ s Web stranice na Web mesto preko povezivača, prikupljajući naslove svih mesta, URL, i najmanje od njihovih tekstualnih sadržaja. Kada nađu mesto, oni pregledaju Web stranice tog mesta i zapisuju sve informacije u indeks.

U slučaju agenata koji pretražuju Web, agentova okolina je Web, a komunikaciju sa korisnikom omogućuje računarski terminal. Ono što agent percipira su reči HTML dokumenta naučene korišćenjem programskih detektora (senzora) povezanih kroz celu mrežu (Internet) uz pomoć HTTP-a. Zadatak agenta je ustanoviti da li je postignut cilj, tj. da li je pronađena Web stranica koja sadrži ključnu reč ili frazu, i ako nije, pronaći druga mesta koja će posetiti i potražiti u svrhu postizanja cilja, tj. obavljanja zadatka. Agent deluje na okolinu koristeći izlazne metode kako bi obavestio korisnika o statusu pretraživanja ili krajnjim rezultatima, koji bi trebali predstavljati postignut cilj.[14]

Rad pretraživača mogao bi se predstaviti na sledeći način. Web agent pretražuje bazu podataka, izlistava reči koje zadovoljavaju upit (ključnu reč) i zapisuje mesto gde ih je našao, indeksira pronađene reči u indeks prema vlastitom sistemu procene, kodira podatke na sigurno mesto i sprema podatke za upotrebu.

[14] www.acm.org/crossroads/xrds5-4/webhunting.html

Indeksiranjem (sastavljanje liste dokumenata na način da sa svake stranice koju pronađe sakupljaju ključne reči i ubacuje ih u indeks reči) se izvlače informacije iz dokumenata i sprema ih u svoju bazu podataka. Cilj je posetiti milione Web mesta i ostati s njima u vezi koliko god je moguće. U svom hodu po mreži razni pretraživači međusobno dele više baza podataka. Što se sprema u bazu podataka, zavisi od određenog pretraživača. Neki indeksiraju svaku reč iz Web dokumenata, a drugi indeksiraju samo naslov. Kada se sprovodi pretraživanje preko ključne reči ili fraze, pretražuje se cela baza podataka, ali u rezultatu se pokazuju samo Web stranice u kojima je nađena ta ključna reč ili fraza. Baza podataka sadrži prikupljene stranice, a pretraživački program na osnovu upita pretražuje bazu podataka, rangira dobijene rezultate i konačno ih šalje korisniku. Uz prikupljanje novih Web stranica agenti se još mogu koristiti i za osvežavanje sadržaja u bazi. Osim potrage za novim stranicama, agenti se u pravilnim vremenskim razmacima, najčešće mesečno, vraćaju na lokacije koje su ranije posetili i proveravaju da li je u međuvremenu sadržaj Web stranica promenjen, i ako jeste, onda izmene dodaju u svoje indekse.[15]

Iako su pretraživački mehanizmi dosta uspešni, ipak se milionima korisnika događa da klikću po linkovima koje dobiju kao rezultat pretraživanja i ne dođu do Web stranice s traženom informacijom. Pretraživački mehanizam često pronalazi ono što korisnik želi, ali ne uvek. Da bi se izbegli ovakvi problemi, razvijeni su posebni programi koji deluju u ime korisnika, te sadrže određeni stepen inteligencije. Ti programi nazivaju se inteligentni agenti za pretraživanje Web-a (eng. Intelligent Web searching agent).

[15] http://prelog.chem.pmf.hr/~tezak/preinin/preinin/index.html

5.1. Inteligentni agenti za pretraživanje Web-a

Izražena dinamičnost i velika distribuiranost podataka zahteva od računarskih programa ne samo da odgovaraju na zahteve, nego da i inteligentno deluju u aktivnostima u svom okruženju, i da se konstantno prilagođavaju trenutnom stanju svoje okoline i na taj način aktivno traže način kako da što bolje ostvare interese korisnika. Važnost informacija je danas ogromna, ali je još važnije prave informacije imati u pravo vreme. Koliko smo samo puta kao rezultat pretraživanja dobili niz podataka koji nisu ispunili naša očekivanja.

Inteligentnim agentima za pretraživanje Web-a nazivaju se kompjuterski programi koji samostalno izvode neki pretraživački posao "u ime i za račun" korisnika. Smešteni su u kompjuteru vlasnika, što ne mora nužno biti (a najčešće i nije) kompjuter krajnjeg korisnika, već neko internet web mesto. Korisnik ih mora "napuniti" informacijama o područjima svoga interesa, pravilima pretraživanja, prioritetima, eventualnim vremenskim ograničenjima. Nakon što agent obavi postavljeni zadatak, analiziraju se rezultati, a agent ispravlja sam sebe ako ti rezultati nisu zadovoljavajući prema nekom unapred postavljenom kriterijumu.[16]

Inteligentni agenti za pretraživanje Web-a mogu imati različite nivoe samostalnosti u izvršavanju zadataka u odnosu prema korisniku i njegovim potrebama. Tako neki od njih mogu "samo" prikupljati informacije, neki drugi filtrirati (selektovati) poruke primljene e-poštom, a neki sarađivati s drugim agentima, te obavljati vrlo složene i specifične zadatke. Razvijene su tri vrste takvih agenata:

- Web puzači (eng. Web crawler),
Pokušavaju korisniku dati celovit pregled informacijske ponude, "šetajući" Web-om i izveštavajući korisnika o onome što su pronašli.

- Web pauci (eng. Web spider),
To su programi koji se pomoću "crvića" (eng. worm) "uvlače" u dubinu hipermedijskih dokumenata, odnosno web stranica, te pronađene informativne sadržaje indeksiraju i pohranjuju u vlastite baze podataka, koju korisnik kasnije može jednostavno lokalno (na svome kompjuteru) pretraživati i analizirati.

- Web roboti (eng. Web robot).
To su programi koji mogu gotovo potpuno nezavisno od korisnika izvršavati kompletne transakcije, poput kupovine na daljinu, rezervacije avionskih karata, novčanih transakcija itd., u skladu s uputsvima koje im je korisnik jednom ranije dao.

[16] Germano Resconi, L. C. Jain, (2004): „Intelligent Agents: theory and applications"

Područje primene inteligentih agenata za pretraživanje je vrlo široko. Tako se inteligentni agenti koriste i za :

- statističke analize - agenti mogu računati prosečan broj dokumenata po serveru, odnos različitih tipova datoteka, prosečan broj Web stranica, broj Web servera na Internetu…

- osvežavanje URL adresa - glavni problem kod održavanja Web stranica je u tome što adrese na druge stranice mogu postati nevažeće, tzv. 'mrtve veze', ako se ta stranica premesti ili čak ukloni. Zbog toga su razvijeni agenti, čiji je zadatak provera ispravnosti adresa.

- mirroring - je postupak preslikavanja sadržaja Web servera na drugi kompjuter s ciljem smanjenja prometa na nekom delu mreže. Agenti se mogu koristiti za uspostavljanje preslikane slike Web stranica, ali ova metoda još nije u potpunosti razvijena tako da kad dođe do promene neke Web stranice, moraju se osvežavati sve preslikane stranice a ne samo one koje su promenjene.

- indeksiranje - prikupljanje ključnih reči iz Web stranica kako bi pretraživanje bilo što lakše i brže. Najčešće se agenti koriste za prikupljanje naslova i nekoliko prvih paragrafa teksta sa stranice.

- pronalaženje podataka - inteligentni Web agenti imaju veliki potencijal u pronalaženju podataka (eng. data mining - proces pronalaženja uzorka u velikoj količini podataka kroz niz pretraživanja), uz to mogu donositi odluke zasnovane na prethodnim pretraživanjima, da bi obavili što kompleksnije pretraživanje.

- kombinovana upotreba - jedan agent može izvršavati i više spomenutih zadataka.[17]

[17] http://en.wikipedia.org/wiki/Intelligent_agent

5.2. Izgradnja jednostavnog inteligentnog agenta za pretraživanje Web-a

Kako izgraditi jednostavnog inteligentnog agenta za pretraživanje Web-a. Parametri koje treba definisati za agenta su:[18]

- cilj – ključna reč ili fraza,
- početna lokacija – URL adresa od koje počinje pretraživanje,
- broj ponavljanja – koliko ponavljanja (što više URL-ova to bolje),
- vremensko ograničenje – koliko dugo će pretraživati,
- metoda pretraživanja – na koji način će birati put.

Razvoj inteligentnog agenta podrazumeva 4 glavne faze:

- Faza inicijalizacije (initialization) uključuje postavljanje svih varijabli, struktura i polja, kao i unošenje ulaznih informacija o pretraživanju: ključna reč, cilj pretraživanja, broj Web stranica koje sadrže ključnu reč (pronađen u ograničenom vremenu), početna lokacija, metoda pretraživanja (ukoliko ih agent podržava više).

- Faza percepcije (perception) je usmerena na korišćenje znanja za kontaktiranje Web stranice i pronalaženje informacija sa te lokacije. Treba identifikovati prisutnost ključne reči i linkove sa te stranice na druge.

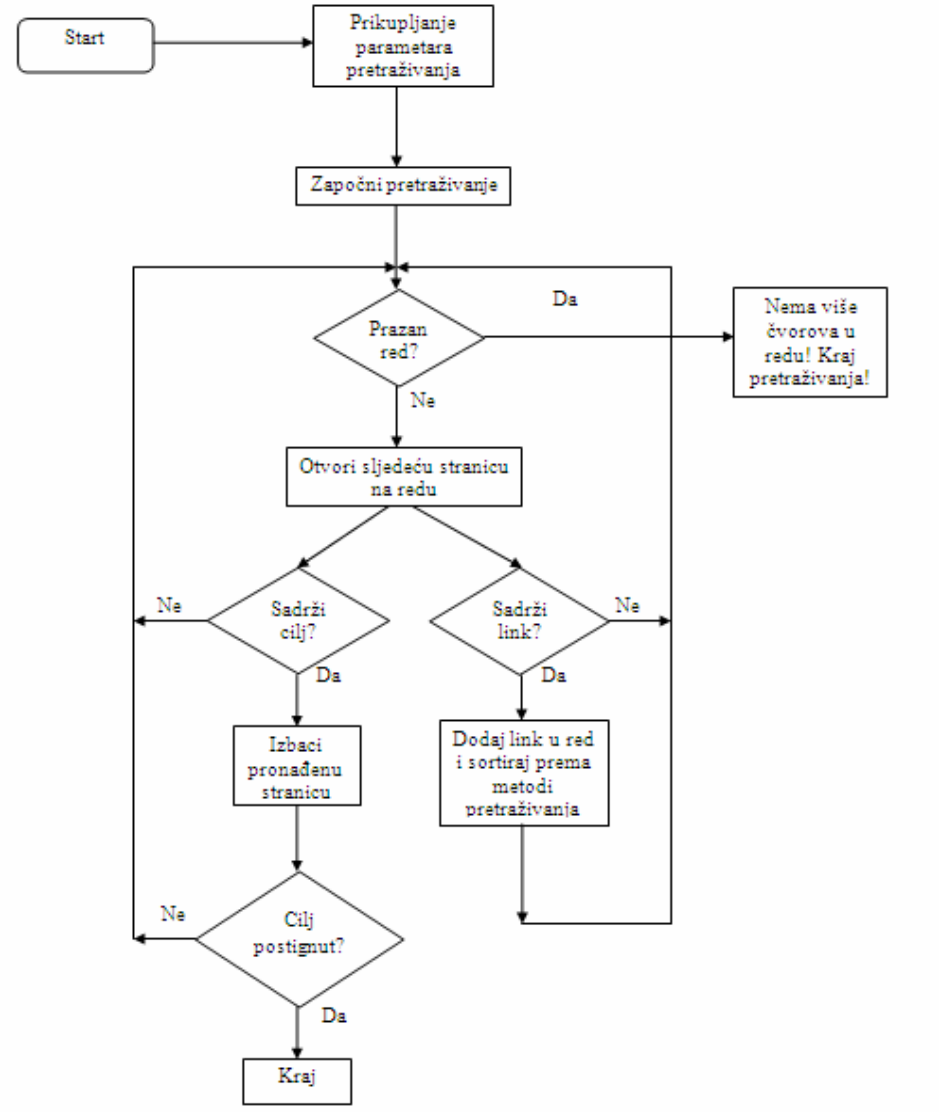
- Faza akcije (action) – tokom ove faze agent uzima sve prikupljene informacije i odlučuje da li je cilj postignut tj. da li je meta pronađena i pretraživanje gotovo. Ukoliko cilj nije postignut mora odlučiti koju adresu će sledeću posetiti. Ovaj deo odnosi se na inteligenciju agenta, a koliko će pametan biti zavisi od metode pretraživanja koju će izabrati.

- Faza efekta (effect) – u ovoj fazi agent izlistava Web adrese na kojima je pronašao ključnu reč, daje status i povratnu informaciju korisniku.

Tokom faze inicijalizacije stvara se struktura podataka koja sadrži informacije o posećenim lokacijama. Za svako pretraživanje moramo jasno definisati cilj i ustanoviti metu. Za početak, radi jednostavnosti, počnemo sa pretraživanjem po ključnoj reči. Korisnik će sigurno želeti da pogleda više od samo jedne pronađene Web stranice s ključnom reči, pa bi broj stranica takođe trebalo da bude specificiran kao cilj. Pošto korisnik želi povratnu informaciju što je brže moguće, treba postići ciljani broj stranica u najkraćem mogućem vremenu. Ukoliko je agent fleksibilan pa dozvoljava odabir metode pretraživanja, treba ustanoviti koju od ponuđenih metoda upotrebiti. Karakteristike pojedinih metoda predstavljaju ponašanje agenta. Program može biti postavljen tako da pita korisnika kakvog agenta želi: agresivnog, umerenog ili prigodnog. Mogu se primeniti i razne kombinacije i permutacije karakteristika pretraživanja. Agent može započeti s pretraživanjem umerenim tempom i, zavisno od rezultata pretraživanja, preći u agresivan mod. Nakon inicijalizacije i definisanja metode pretraživanja, počinje samo pretraživanje.

[18] http://www.cs.berkeley.edu/~russell/aima1e/chapter02.pdf

Web agent ulazi u fazu percepcije otvaranjem komunikacionog mehanizma na HTTP portu za specificiranu Web stranicu. Nakon uspostavljanja veze, HTTP naredba GET zajedno s informacijom o direktorijumu poziva server da ispostavi zatraženu stranicu. Tada se dobijena stranica gramatički analizira radi pronalaženja linkova, koji se onda poredjaju u čvorove kako bi se moglo ustanoviti da li je prisutna tražena meta. Ključni HTML tagovi koji ukazuju na linkove su "HREF=" i "SRC=", a link u sebi mora sadržati "http" negde u tekstu. Ukoliko tražimo neki drugi tip podataka možemo napraviti takvog agenta da traži npr. slike ("JPG" ili "GIF" reference) ili bilo koji tip dokumenta koji se lako prepoznaje. Mora se paziti da se od agenta ne zahtevaju tipovi dokumenata kojima on ne zna da rukuje. Web stranice mogu sadržati različit broj hyperlinkova ili referentnih HTML stranica - počevši od nijednog pa do stotina ili čak hiljada. Pametno bi bilo postaviti gornju granicu broja linija koje će agent primiti. Gramatička analiza teksta može biti programski zadatak pogotovo ako se radi u C-u, ali korišćenjem napredne biblioteke ili nekih oblika jezika za gramatičku analizu teksta, zadatak se može znatno pojednostaviti.



Slika 4: Niz osnovnih akcija inteligentnog Web agenta

Nakon što je stranica prihvaćena i klasifikovana, sledi faza akcije. Kako je agent uzeo Web stranicu i izvadio iz nje važne informacije, sada je vreme za malo "mišljenja" i akcije. Ako je cilj postignut, tj. pronađen je traženi broj Web stranica onda je pretraživanje završeno i bilo je uspešno. Ukoliko traženi broj Web stranica nije postignut, sledi donošenje odluke koji je sledeći korak. Središte odluke je izbor algoritma pretraživanja. U red se stavljaju čvorovi (linkovi) iz tekućeg dokumenta. Strategija pretraživanja određuje kojim redosledom će se oni posećivati i na koji način. Što je strategija pretraživanja informisanija, to je veća verovatnoća da će cilj biti postignut. Slepa pretraživanja obično ne mare za dodatne informacije, jedino mogu zatražiti informaciju o ceni koja je za Web stranice određena kao njihova udaljenost od Web agenta. Udaljenost, ili brzina pristupa, može se koristiti kao heuristička informacija budući da se svaki čvor može posmatrati kao deo cilja, a deo cilja je umanjiti ukupno vreme pretraživanja. Osim brzine pristupa heuristička informacija može biti i zemljopisna udaljenost tj. zemlja kojoj Web stranica pripada. Tako je npr. većina Web stranica u SAD-u u domenama "COM", "NET", "EDU", "ORG", "GOV", "MIL" ili "US" pa agent iz SAD-a može stranicama koje nisu u ovim domenama dodeliti veću cenu jer nisu iz njegove zemlje i verovatno su na drugom jeziku. Ova heuristika je prilično slaba jer mnoge institucije izvan SAD-a imaju "COM" ili "EDU" adrese (kompanije ili obrazovne institucije). Prema dogovoru kako se dodeljuju IP adrese i imena Web stranicama, ne postoji sasvim tačan način kako odrediti lokaciju Web stranice. Neznatno olakšanje snalaženja u ovom haosu daje CIDR (Classless Inter-Domain Routing) zapis IP adrese, odnosno besklasno adresiranje koje uz običnu adresu nosi informaciju o IP prefiksu koji predstavlja broj bitova koji pripadaju mrežnom delu adrese (mrežna maska), ali s obzirom na broj IP adresa, pomoć je zaista minimalna.

Nakon što metoda pretraživanja doda ili izbriše čvorove može ih sortirati po ceni i/ili heuristici. Izbor algoritma sortiranja nije toliko važan, ali naravno da će onaj koji radi brzo i tačno poboljšati performanse agenta i skratiti vreme pretraživanja. Ukoliko je prostor pretraživanja iscrpljen i nema više čvorova u redu, agent može dobiti novu početnu vrednost i početi odatle ili poslati status i završiti s radom. U fazi efekta Web agent izveštava korisnika o statusu pretraživanja i pokazuje mu rezultate. Efektor iz akcije zasnovan na percepciji okoline mora dati povratnu informaciju korisniku. Web agent se premešta iz inicijalne faze u petlju koja se sastoji od faza percepcije, akcije i efekta, sve dok se cilj pretraživanja ne postigne ili dok se ustanovi da se ne može pronaći traženo.

6. Metode pretraživanja web-a

U području veštačke inteligencije, rešavanje problema se može okaraktersati kao sistematično pretraživanje mogućih ishoda u svrhu pronalaska rešenja. Rešavanje problema putem algoritama za pretraživanje je vrlo česta tehnika. U kompjuterskoj znanosti algoritam za pretraživanje u širem kontekstu je algoritam koji daje rešenje nakon vrednovanja skupa mogućih rešenja. Skup svih mogućih rešenja problema naziva se prostor pretraživanja.

Zadatak agenta je doći iz početnog stanja do ciljnog stanja, koje je rešenje problema, sledom događaja koji uzrokuju promene stanja. Čest je slučaj da skup stanja između početnog i ciljnog nije jednoznačno određen, nego postoji više različitih, često i različito efikasnih puteva od početnog do krajnjeg stanja. Za rad takvih sistema potrebno je ugraditi algoritme za traženje spomenutih puteva. Otežavajuća okolnost kod izrade spomenutih algoritama je nepostojanje formalnih postupaka koji bi na temelju početnog i krajnjeg stanja sistema određivali skup stanja na putu između njih. U takvim slučajevima potrebno je pribeći metodi pretraživanja prostora stanja.

Pretraživanje prostora stanja je postupak koji sledno prolazi kroz stanja u kojima bi se mogao nalaziti sistem, pa upoređivanjem trenutnog stanja s ciljnim stanjem utvrđuje da li je postupak došao do kraja. Skup stanja organizovan je u strukturu grafa ili stabla pa se pretraživanje svodi na pronalaženje puta kroz graf ili stablo. Čvorovi grafa pritom predstavljaju stanja sistema, a dva su čvora povezana granom u grafu ukoliko se iz jednog stanja može preći u drugo. Osim toga, svaki čvor ima vezu na čvor prethodnik tako da se može pronaći put u grafu kad se pronađe konačno stanje. Put se gradi tako što se krene od ciljnog čvora, pa se sledeći vezuje na prethodnike, dok ne dođe do početnog čvora grafa.

Problem, može imati i složeniju strukturu. Postoji celi izbor mogućih događaja od kojih niko ne garantuje da će nas upravo taj dovesti do rešenja. Opšti pristup rešavanja takvog procesa je generisati i ispitati tj. primeniti neku akciju kako bi proizveli novo stanje, a zatim ispitati da li je to stanje ciljno stanje i ako nije ponoviti proceduru. Procedura ponavljanja procesa generisanja i ispitavanja je sam proces pretraživanja. Pretraživanje može biti kratko ili dugo, zavisno od složenosti problema i efikasnosti agentove strategije rešavanja problema. U nekim slučajevima određeno rešenje je bolje od drugog rešenja iako oba predstavljaju konačno rešenje, tj. cilj. Tada se uvodi pojam vrednost putanje koja predstavlja zbir vrednosti pojedinih koraka putanja.[19]

Agent generiše sve moguće ishode nekog događaja, bira one koji mu odgovaraju za pronalaženje navedenog cilja i izvršava niz koraka počevši od početnog (inicijalnog) stanja kako bi stigao do cilja. U slučaju inteligentnog agenta za pretraživanje radi se o korišćenju algoritama pretraživanja za "navigaciju" kroz World Wide Web kako bi agent došao do cilja, tj. pronašao ono što mu je zadato.

[19] http://pespmc1.vub.ac.be/PROBSOLV.html

Razlikujemo dve osnovne klase algoritama za pretraživanje: slepo ili neusmereno pretraživanje (eng. blind search or uninformed search) i heurističko ili usmereno pretraživanje (eng. heuristic or informed search).[20]

1) Slepo pretraživanje je pretraživanje koje nema nikakvih informacija o broju koraka ili vrednosti putanje (vrednost udaljenosti čvora od početnog čvora) od početnog do krajnjeg stanja, tj. cilja. U ova pretraživanja spadaju:
− pretraživanja po dubini (eng. Depth-first search)
− pretraživanja po širini (eng. Breadth-first search)
− pretraživanja s jednolikom cenom (eng. Uniform-cost search)
− pretraživanje do određene dubine (eng. Depth-limiting search)
− učestalo pretraživanje po dubini (eng. Iterative deeping search)

2) Heurističko pretraživanje je pretraživanje koje ima dodatne informacije o cilju, cenu putanje ili broj koraka. Te informacije čine ova pretraživanja boljim od slepih i tako im omogućavaju gotovo racionalno ponašanje. U ova pretraživanja spadaju:
− pretraživanje najboljim prvim (eng. Best first search)
− pretraživanje penjanjem (eng. Hill-climbing search)
− A* pretraživanje (eng. A* search)
− ograničeno pretraživanje po širini (eng. Beam search)
− IDA* pretraživanje (eng. Iterative deeping A* search)

Metode pretraživanja razmatraju se u okviru ovih kriterijuma:
- Potpunost: da li strategija pretraživanja garantuje pronalazak rešenja?
- Vremenska složenost: koliko je vremena potrebno za pronalazak rešenja?
- Složenost prostora: koliko je memorije potrebno za pretraživanje?
- Optimalnost: da li strategija pretraživanja pronalazi visoko kvalitetno rešenje među više rešenja?

Vremenska i prostorna složenost se mere u terminima:
- b: najveći faktor grananja stabla pretraživanja
- d: dubina najkraćeg puta do rešenja (koren je dubine 0)
- m: najveća dubina stabla pretraživanja (može biti beskonačno)[21]

[20] www.acm.org/crossroads/xrds5-4/webhunting.html
[21] http://www.course.cs.york.ac.uk/lpa/slides2006/uninformed-search.pdf

6.1. KQML

KQML (Knowledge Query and Manipulation Language) predstavlja jezik i skup protokola za komunikaciju inteligentnih agenata. Podrazumeva raznovrsne arhitekture i tipove inteligentnih agenata, kao i komunikaciju agenata sa drugim agentima i serverima. KQML koristi standardne protokole za razmenu informacija kao što su TCP/IP, email, HTTP, CORBA. On tretira sve procese koji komuniciraju kao agenti zasnovane na znanju. Time je komunikacija agenata sa drugim procesima olakšana. Ovde se za modeliranje agenata sa znanjem koriste dve vrste znanja i to:

- Baza činjenica
- Baza ciljeva

Da bi agenti komunicirali razmenom poruka moraju da se usaglase. KQML ne zadaje nikakva ograničenja pa je i upravo zbog toga dobar.[22]

[22] http://en.wikipedia.org/wiki/Knowledge_Query_and_Manipulation_Language

7. Google

Pojam „Google“ bazira se na igri reči (neki izvori tvrde da se radi o pravopisnoj grešci) kod američkog izgovora reči googol. Milton Sirotta (nećak američkog matematičara Edwarda Kasnera) navedeni je pojam izmislio 1983. godine, kako bi broju sa jedinicom i sto nula dao naziv. Osnivači Google-a slično tome bili su u potrazi za pojmom koji bi označio superlativnu količinu informacija koju bi njihov pretraživač trebao pronaći.

„Google“ je naziv poznatog internet pretraživača koji momentalno dominira informatičkim tržištem. Probna verzija pretraživača „Google“ pokrenuta je 7. septembra 1998. Izgled Google-a se od tada praktično nije promenio. Zahvaljujući uspehu pretraživača, kompaniji Google Inc. omogućeno je finansiranje daljnjih programa; pristup njima omogućiće se preko službenih stranica Google-a. Usprkos svemu, jezgro Google-a je internet pretraživač. Takođe je vredno napomenuti da konkurent Yahoo povremeno koristi Google-ove baze podataka u vlastite potrebe.



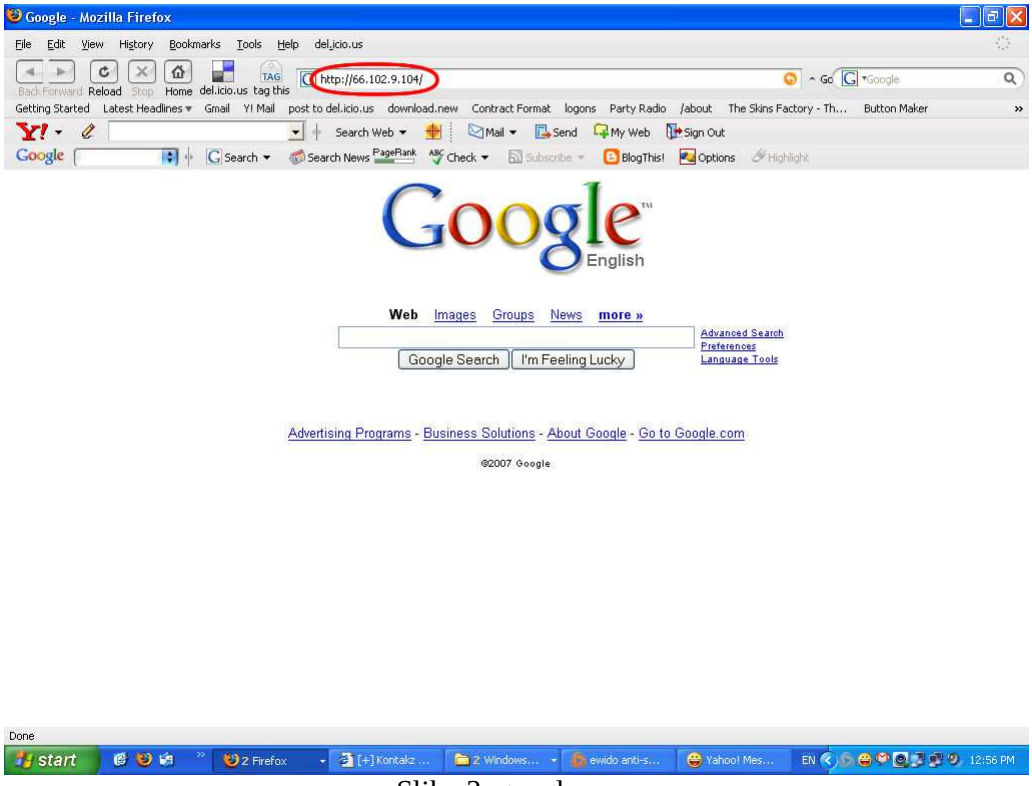
Slika 2. Google pretraživač

Uspeh Google-a kao kompanije započeo je uspehom njihovog pretraživača. Dok su drugi pretraživači korisnike mučili komplikovanim upitima, lošim rezultatima i veoma dugim vremenom odgovora, Google je svoje korisnike do rezultata dovodio na lak i jednostavan način. Uprkos svemu; da bi se moglo uspešno koristiti Googleovim mogućnostima potrebno je i ponešto znati.

Za razliku od ostalih „dot com“ kompanija, Google karakteriše veoma uspešan, konstantan i progresivan stil razvoja. Kad je IT tržište, pogotovo područje interneta (world wide web) bilo zasićeno; bolje reći prezasićeno, došlo je do katastrofe neizmernih razmera. Brojne kompanije i pojedinci finansijski su uništeni. Paralelno s tim, Google se razvija bez većih poteškoća; čak bi se moglo reći da raste neslućenom brzinom.

Zašto je Google uspešan u isto vreme kad se konkurencija bori za život? Kao odgovor na to pitanje mogli bismo navesti neki oblik prirodne ravnoteže. Stari, praktično tromi sistemi zamenjuju se novim; tj. dinamičnim i inovativnim sistemima. Ključna reč navedenog razdoblja je inovacija. Glavna karakteristika Google-a i uzrok njegovog uspeha su neprestane inovacije. Većina usluga i produkta na kojima se temelji Google plod su inovacija i nekonvencionalnog pristupa rešavanju problema.

Bazirajući se na već navedenim tržišnim načelima Google je postao kompanija koja doslovno dominira internetom. Kad se spominju internet pretraživači, programi oglašavanja i distribucije medijskog sadržaja internetom, Google ili barem neka kompanija ili manja kompanija u vlasništvu Google-a, ne samo da posluju u određenom sektoru, već su najčešće neophodni sastavni deo tog sektora. Internet se danas praktično ne može zamisliti bez Google-a. Mogli bismo reći da internet bez Google-a nije u stanju da funkcionise na način koji danas poznajemo i koristimo.[23]



Slika 3. google.com

[23] http://en.wikipedia.org/wiki/Google

7.1. Pojam „googlati“

Google je kao pretraživač postao toliko popularan da su određeni pravopisi počeli koristiti pojam „googlati“. Navedeni glagol više ne predstavlja samo pretraživanje pomoću pretraživača Google, već je sinonim za pretragu interneta bilo kojim pretraživačem. Od avgusta 2006. Google želi da pravno zaštiti pojam „googlati“, tako da bi se navedeni pojam smeo koristiti samo u vezi sa interenet pretraživačem „Google“. Google želi da spreči razvoj događaja koji se odvijao sa pojmovima kao što su npr. Tempo, Nutella, Jeep, Linoleum, Aspirin, Gramofon, itd. – navedeni pojmovi pre su predstavljali zaštićene logotipe, a danas su sinonimi za razne proizvode.

U Australiji su već 2002. Sony-u oduzeta marketinška prava za pojam „Walkman“, jer se navedeni naziv koristio za svaku vrstu prenosivog kasetofona. Advokati Google-a insistiraju na tome da sve redakcije novina i časopisa moraju momentalno prestati da koriste pojam „googlati“ u smislu „pretraživanja interneta“.[24]

[24] http://radicalproject.blogspot.com/2011/06/googlati-nekoga.html

7.2. Arhitektura

Google Inc. širom sveta poseduje više kompjuterskih centara, koji pojedinačno sadrže potpunu funkcionalnost pretraživača. Korisnički upit se kroz „Domain Name System“ (sistem koji imena kao što je google.com pretvara u IP-Adrese) prenosi u idealnom slučaju u sedeći mrežno topološki najbliži cenatar – navedeni centar je samo u retkim slučajevima identičan najbližem geografski topološkom centru. U slučaju ako određeni kompjuterski centar nije sposoban za rad, ostali centri preuzimaju njegovu ulogu.

Svaki kompjuterski centar sastoji se od računarskog clustera. Za tu namenu korišteni kompjuteri su personalni kompjuteri kompatibilni s IBM-ovim proizvodima, prvenstveno kompjuterima; tj. sastoje se od relativno jeftinih standardnih komponenti. Navedeni sistem koristi samorazvijeni Google File System – razdeljena arhitektura koja sve podatke suvišno pohranjuje na različitim uređajima. U slučaju ako dođe do kvara određenog kompjutera ili određenog hard diska unutar kompjutera, odgovarajući podaci se s drugog mesta unutar clustera prekopiraju na zamenski kompjuter, tako da se oštećena hardware-ska komponenta može zameniti u toku rada, bez zaustavljanja sistema. Celokupni sistem može se veoma lako modifikovati, tj. prema potrebi se mogu dodavati ili oduzimati kompjuteri ili pojedinačni hard diskovi.

Odlaganjem određene strukture podataka (indexi, kopije dokumenata...) na različitim kompjuterima i zahvaljujući velikom broju raspoloživih kompjutera pojedini upiti mogu se relativno brzo obraditi, budući da se informacije čitaju paralelno. Vreme odgovora od maksimalno pola sekunde smatra se prihvatljivim, te se koristi za orijentaciju kod utvrđivanja brzine kojom se isporučuju odgovori na upite. Brza „dostava“ informacija bila je veoma velik doprinos popularnosti Googlea već početkom razvoja sistema. Konkurenciji je za odgovor trebalo mnogo više vremena, posebno kod inkrementnih upita, tj. kod detaljnih upita korisnika ušteda vremena znatno je veća.

7.3. Redosled stranica

Broj dokumenata koji odgovaraju upitu može biti nepregledne veličine. U tom slučaju potrebno je rezultate sortirati prema relevantnosti, kako bi se na upit odgovorilo s najvažnijim dokumentima pozicioniranim na prvim mestima. Procena upotrebljivosti izvora u odnosu na upit jedan je od odlučujucih kriterijuma prema kojima se meri kvalitet pretraživača. Na temelju količine za to određenih informacija dodeljeni Ranking se s vremena na vreme automatski proverava pomoću algoritma. Proces rangiranja stranica mora se obaviti u razumnom vremenskom okviru, kako bi Ranking bio aktuelan.

Google za organizaciju traženih rezultata među ostalim koristi i tzv. PageRank vrednost. PageRank se određuje prema broju linkova na određeni dokument. Koliko više postoji linkova, i koliko je veći Page Rank stranice na kojoj su smešteni linkovi, toliko je veća vrednost odgovarajućeg dokumenta, tj, dokumenta na koji ukazuju linkovi. Pojedinačni PageRank stranice može se odrediti pomoću browser-plugina pod nazivom Google-Toolbar. Jedini problem je u tomu što je navedeni PageRank najčešće star par meseci. Takođe postoje i online službe koje omogućuju pregled PageRanka određene stranice.

Uz PageRank, kod redosleda stranica u obzir se uzimaju i daljnji faktori, kao što je zastupljenost ključnih reči unutar teksta ili naslova dokumenta. Takođe veliku ulogu imaju i imena sidra „named anchor“ unutar dokumenta. Poslednja tehnika se veoma često zloupotrebljava kod izrade tzv. „Google bombe“. Tačan način orgnizacije Googleovih rezultata je poslovna tajna, kako bi se sprečila manipulacija rezultata pretraživača. Optimizacija stranica za pretraživače bavi se navedenim metodama s ciljem da postigne što bolje rezultate uz što manje troškove za vlasnika stranice.

7.4. Veličina indexa i dodatne činjenice o Google-u

Okviran broj dokumenata u indexu pretraživača za vreme prvih sedam godina prikazan je ispod:

Osnivanje poduzeća, Januar 1998: 25.000.000
August 2000: 1.060.000.000
Januar 2002: 2.073.000.000
Februar 2003: 3.083.000.000
Septembar 2004: 4.285.000.000
Novembar 2005: 8.058.044.651

Osim toga, u Junu 2005. prema izjavi Google-a postojao je index od 1.187.630.000 fotografija, milijarda usenet članaka, 6.600 kataloga i oko 4500 novinskih izvora. Od sedmog rođendana kompanije u Septembru 2005. veličina indexa se više ne prikazuje na glavnoj stranici. Prema određenim stručnjacima detaljno numerisanje stranica unutar Googlevog indexa ne postoji. Google momentalno na svojim stranicama spominje index veličine „više od 8 milijardi web stranica“.

Google-ova tehnika pretraživača sa licencom se dodeljuje raznim internet portalima, kako bi se i tamo omogućilo pretraživanje interneta bez toga da korisnik napušta portal.

„Google Appliance“ je oblik u kojem Google prodaje vlastitu tehniku pretraživanja kompanija koje žele koristiti u intranetu. Kod Appliancea se radi o serveru sa predinstaliranim softwareom, koji u mreži firme preuzima ulogu koju inače Google ima unutar World Wide Weba.

Kao operativni sistem koristi se „osiromašena“ verzija Linux distribucije Red Hat. Programski jezik C++ i Python imaju esencijalnu ulogu kod implementacije softwarea pretraživača u celokupni sistem.

Dok je Google bio još samo istraživački projekat, 1996. godine, za smeštaj podataka mu je bilo dovoljno svega 40 GB. Kako u to vreme hard diskovi tog kapaciteta nisu postojali, za to im je trebalo 10 diskova od po 4GB. Da bi se oni lako mogli menjati, bilo im je potrebno naći neko praktično kućište koje ne košta puno. Tada je jedan od inženjera odlučio da od lego kockica napravi "skladište" odnosno bolje reći postolje za deset ovih uredjaja. Ovo se tada pokazalo kao vrlo praktično rešenje, a kako "urbane" legende kažu, raspored boja na kockicama je imao presudan uticaj na boju logoa firme. Naravno ovo nikada nije zvanično potvrdjeno.[25]

[25] http://www.zezas.me/2010/12/10-cinjenica-o-guglu-koje-verovatno.html?utm_source=linkeri.com&utm_medium=referral&utm_campaign=linkeri

7.5. Upotreba

Uz jednostavnu pretragu, Google nudi i mogućnost proširene pretrage, tj. primenu određenih filtera.

Jednostavna pretraga kod unosa standardno koristi boolski operator „i“ (Konjunkcija: „i“ ili „+“ kod unosa više pojmova). Kod tog procesa traže se svi dokumenti u kojima se spominju traženi pojmovi. Uz PageRank kod navedenog procesa rezultati se boduju i prema lokaciji ključnih reči, tj. nalaze li se one u naslovu stranice, imenu dokumenta, zapisu, natpisu ili običnom tekstu.

Uz boolsku standardnu vezu „i“ može se koristiti i boolski „ili“ (Disjunkcija OR „|“). Primer: „plaža OR beach“. Google kod toga obraća paznju na sve reči koje se nalaze unutar OR lanca. U slučaju ako je na raspolaganju više reči, one se povezuju sledećim principom: „reč 1 OR reč 2 OR reč 3 ...“. Kod tražilice funkcioniše samo sledeći normativni oblik, umesto npr. (voziti auto) OR (leteti avionom) mora se uneti sledeći oblik (auto OR leteti) (avion OR voziti) (voziti OR leteti).

Da bi se pretraga ograničila na manji broj pojmova, pretragu je moguće obaviti po egzaktnim frazama. Traženi pojmovi se tada ne pojavljuju bez reda u dokumentu, već se moraju pojaviti u tačno definisanom redosledu kako bi se dokument u kojem se nalazi prikazao među rezultatima. Fraze se definišu navodnicima, npr. „kompanija Google“. Ne traže se dokumenti u kojima se negde odvojeno nalaze pojmovi kompanija i Google. Traže se samo dokumenta u kojima se nalazi sintagma „kompanija Google“. Po želji se umesto reči može koristiti i *. Pomoću * u tom slučaju, samo se osigurava prostor za reč koja će se definisati prema kriterijumima koje određuje Google pomoću informacija koje je dobio od korisnika.

Kao dalji kriterijum isključivanja koristi se minus „-“. Npr. „Hannibal -Lecter“. Navedeni kriterijum traži vojskovođu iz Antike i istovremeno isključuje filmski lik imena „Lecter“.

Svim navedenim funkcijama može se pristupiti u obliku formulara linkom „napredno pretraživanje“. Takođe se zahvaljujući navedenom formularu mogu koristiti još preciznije definisani kriterijumi.

Često korištena pretraga pomoću tzv.“Wildcarda“ unutar Googleovog indexa nije moguća. Tako se npr. kod pretrage za pojmom „Googl“ ne traže pojmovi „Googlin“ ili „Googler“. Pretragu je moguće obaviti s maksinalno 32 pojma (u obliku pojedinačnih pojmova ili fraze). Navedena opcija je korisna za traženje raznih modifikacija određenog teksta. Zahvaljujući tim opcijama moguće je kreiranje veoma preciznih upita. Rezultati upita su najčešće bolji, ukoliko se postavi relativno inteligentan upit, ukoliko upit nema smisla, ili je banalan, rezultati su najčešće neupotrebljivi.

7.6. Proširena upotreba

Uz jednostavnu sintaksu pretrage može se koristiti i proširena sintaksa pomoću formulara za pretraživanje, tj. opcije „napredno pretraživanje“ Sledeći operatori su nam poznati, operatori se označavaju dvotačkom „ : “.

cache: - Izlistava samo stranice koje je Google memorisao. To može biti korisno kod stranica koje često menjaju sadržaj, kao što su portali. Link koji je katalogiziran kod Googlea stoga ponekad vodi do stranice čiji se sadržaj u međuvremenu promenio ili čak uklonio. Sadržaj je usprkos svemu još određeno vreme memorisan unutar Google-ovog kompjuterskog sistema, pa se tamo može pregledati.

define: - Pretražuje definicije pojmova unutar internet enciklopedija, kao što je Wikipedia i sl.

filetype: ili ext: - Traži dokumente sa određenim naslovima datoteka. Npr. „proračun filetype:xls. Teoretski su mogući svi završeci datoteka, ali Google trenutno može „pregledati“ sledeće formate: PS, PDF, AI, DOC, PPT, XLS, SWF, kao i sve tekstualne datoteke, kao što su RTF, TXT, ASP, PHP, CGI, HTML, LOG, INI, JS, itd..

inanchor: -Traži pojmove samo unutar linkova sidra. Često link sa precizno definisanim nazivom vodi preciznije do cilja, nego da se pojam pojavi samo negde unutar teksta.

inurl: - Izlistava stranice kod kojih se traženi pojam nalazi unutar URL-a.

intitle: - Traže se dokumenti u kojima se traženi pojam ili pojmovi nalaze samo u naslovu datoteke ili stranice.

intext: - Traže se dokumenti u kojima se traženi pojam ili pojmovi nalaze unutar teksta.

link: - Izlistava sve stranice koje poseduju link na određenu stranicu. Primer: link:wikipedia.org . Ovaj operator je važan za optimizaciju Page Ranka i pronalaženje „prijateljskih“ stranica.

site: - Pomoću ovog operatora pretraga se može ograničiti samo na određen domen, što može biti veoma praktično ukoliko stranica ne poseduje vlastitu funkciju pretraživača (npr. ribonukleinska kiselina site:hr.wikipedia.org) ili za brzu pretragu top-level-domene.

related: - Traži slične stranice.

Primenjeni operatori za pretragu mogu se spremiti, tj. Ponoviti i povezati linkovima i u obliku URL-a koji Google ispisuje prilikom pretraživanja interneta. Dalje mogućnosti nude dodatne Google-funkcije.

7.7. Google Chrome

Ne moraju da se navode lični podaci da bi se preuzeo i koristio Google Chrome. Kada se preuzme Google Chrome ili se koristi da biste kontaktirali Google servere, Google prima samo standardne podatke za evidenciju, uključujući IP adresu računara i jedan ili više kolačića. Google Chrome može da se konfiguriše tako da ne šalje kolačiće Google-u ni drugim sajtovima. Osim toga, neke Google Chrome funkcije šalju određene dodatne informacije Google-u.

Kada se u traku za adresu unesu URL adrese ili upit, slova koja se unose se šalju Google-u kako bi funkcija Suggest mogla automatski da preporuči termine ili URL adrese koje se eventualno traže. Ukoliko se odluči da se deli statistika korišćenja sa Google-om i prihvati neki predloženi upit ili URL adresa, Google Chrome će te informacije takođe poslati Google-u.

Ukoliko se unese nepostojeća URL adresa, Google Chrome će je poslati Google-u kako bi mogao da pomogne da se pronađe adresa koja se traži. Ova funkicja se moze onemogućiti.

Funkcija sigurnog pregledanja Google Chrome pretraživača povremeno kontaktira Google servere radi preuzimanja najnovije liste poznatih sajtova sa „pecanjem“ i malverom. Kada se poseti neki sajt za koji se sumnja da se bavi „pecanjem“ ili slanjem malvera, pretraživač će Google-u poslati keširani, delimični primerak URL adrese tog sajta kako bi mogao da pošalje više informacija o toj rizičnoj URL adresi. Google ne može isključivo na osnovu ovih informacija da utvrdi stvarnu URL adresu koja je posećena. Osim toga, ako se izabre da se dele statistički podaci o korišćenju sa Google-om i ako se poseti sajt za koji se sumnja da koristi „pecanje“ ili šalje malver, sa Google-om će se deliti i neki drugi podaci. Oni obuhvataju punu URL adresu stranice koja je posećena, zaglavlje „upućivača“ koje je poslato na tu stranicu, kao i URL adresu koja se podudara sa listom malvera u okviru Google funkcije bezbednog pregledanja. Može da se onemogući Google usluga bezbednog pregledanja ili da se onemogući deljenje statistike o korišćenju sa Google-om .

Primerak Google Chrome pretraživača sadrži privremeni, nasumično generisani broj instalacije koji će biti poslat Google-u kada se instalira i prvi put koristi Google Chrome. Privremeni broj će brzo biti izbrisan kada Google Chrome automatski proveri da li postoje ažuriranja.

Može da se izabere Google kao pretraživač koristeći Google Chrome, a može i da se koristi Google Chrome da bi se pristupilo drugim Google uslugama kao što je Gmail. Kada se pristupa Gmail-u ili drugim uslugama primenjuje se Politika privatnosti odgovarajuće usluge, bez obzira na to koji se pretraživač koristi. Ukoliko se koristi Google Chrome za povezivanje na Google usluge, Google neće dobiti nikakve posebne, niti dodatne lične informacije.

Ukoliko se Google Chrome-u omogući funkcija sinhronizacije, Google Chrome će sačuvati informacije o podešavanjima pretraživača, kao što su podaci o obeleživačima, na Google-ovim serverima povezane s vašim Google nalogom. Informacije sačuvane na nalogu zaštićene su u skladu sa Google politikom privatnosti. Ove informacije se moraju čuvati kako bi bile poslate na druge računare na kojima je Google Chrome funkcija sinhronizacije omogućena.

Ukoliko se koristi funkcija Prevodilac u Google Chrome pretraživaču, Google Chrome će poslati tekst koji se izabre za prevod Google-u kako bi bio preveden.

Ukoliko se koristi funkcija lokacije Google Chrome-a koja nam omogućava da delimo svoju lokaciju sa bilo kojim sajtom, Google Chrome će poslati informacije o lokalnoj mreži Google uslugama lociranja da bi dobio procenjenu lokaciju. U zavisnosti od mogućnosti uređaja, informacije o lokalnoj mreži uključuju informacije o bežičnim ruterima koji su najbliži, ID-ovima mobilnih telefona predajnika mobilne mreže koji su najbliži, jačinu bežičnog ili mobilnog signala i informacije o uređaju kao što je IP adresa uređaja. Ove informacije se koriste za obradu zahteva lokacije i za rad, podršku i poboljšanje sveukupnog kvaliteta Google Chrome-a i Google usluga lociranja. Gore navedene informacije koje se prikupe biće anonimne i objedinjene pre nego što ih Google upotrebi za razvoj novih funkcija ili proizvoda i usluga ili za poboljšanje sveukupnog kvaliteta bilo kojih Google-ovih proizvoda ili usluga.

Može da se izabere da se Google-u šalje statistika korišćenja i izveštaji o otkazivanju programa. Statistika korišćenja i izveštaji o otkazivanju programa pomažu da se utvrde problemi sa pretraživačem, razume kako korisnici stupaju u interakciju sa pretraživačem i poboljšavaju njegove performanse. Google Chrome pokušava da izbegne slanje informacija koje nas identifikuju. Međutim, izveštaji o otkazivanju programa mogu da sadrže podatke iz datoteka, aplikacija i usluga pokrenutih u trenutku kada je došlo do problema u radu. Može da sa trećim stranama deli izvesne, objedinjene, opšte podatke koji se dobiju analizom, kao što je učestalost određenih vrsta otkazivanja programa.[26]

[26] http://en.wikipedia.org/wiki/Google_Chrome

7.8. Podaci koji se čuvaju na računaru prilikom korišćenja Google Chrome

Google Chrome evidentira korisne podatke o istoriji pregledanja na računaru. Čuvaju se sledeće informacije:

- osnovni podaci o istoriji pregledanja: URL adrese stranica koje su posećene, keširana datoteka teksta sa tih stranica i lista nekih IP adresa do kojih vode veze sa stranica koje su posećene;
- pretraživi indeks većine stranica koje su posećene (izuzev sigurnih stranica sa „https" web adresama, kao što su stranice nekih banaka);
- snimci ekrana većine stranica koje su posećene u vidu sličica;
- kolačići koje na računaru ostavljaju web sajtovi koji su posećeni;
- evidencija preuzetih stavki sa web sajtova.

U svakom trenutku može da se izbriše cela istorija ili samo određeni delovi.

Može i da se ograniči koje podatke Google Chrome čuva na računaru u režimu bez arhiviranja. U ovom režimu, pretraživač neće čuvati osnovne podatke o istoriji pregledanja kao što su URL stranice, keširani tekst stranica ili IP adrese stranica do kojih vode veze sa web sajtova koji su posećeni. Pretraživač neće čuvati ni snimke stranica niti evidentirati preuzete stavke. (Ali, ovi podaci mogu da se čuvaju na nekom drugom mestu na računaru.) Novi kolačići koji su dobijeni u režimu bez arhiviranja neće biti sačuvani kada se zatvori pretraživač ili se vrati u normalan režim pregledanja. Može da se vidi kada je korisnik u režimu „Bez arhiviranja" po tome što će se ikona režima „Bez arhiviranja" pojaviti u gornjem uglu pretraživača; ponekad se promeni i boja okvira prozora pretraživača.

Čuvaju se i podaci o izmenama konfiguracije pretraživača, poput obeležavanja web stranice ili promene podešavanja. Režim bez arhiviranja ne utiče na ove promene. Može da se odabere da Google Chrome čuva lozinke za određene web sajtove. Sačuvane lozinke mogu da se pregledaju na kartici „Lične stvari" u dijalogu „Opcije".

7.9. Informacije o korišćenju dodataka za Google Chrome pretraživač

Ukoliko se instalira dodatak za Google Chrome, on može da čuva podatke lokalno i šalje sve podatke kojima ima dozvoljen pristup na bilo koji server nezavisnog proizvođača sa kojim mu je dozvoljena komunikacija. Kada se preuzme dodatak, dobija se i obaveštenje o dozvolama koje se daju dodatku. Google Chrome na računaru čuva listu svih dodataka, zajedno sa URL adresom za ažuriranja i kategorijom dozvola potrebnih za funkcionisanje dodatka. Google Chrome povremeno koristi ove informacije da bi proverio da li ima ažuriranja za dodatke i da bi automatski preuzeo i instalirao ažuriranja. Pored toga, Google Chrome čuva listu dodataka za koje je utvrđeno da su štetni ili nelegalni, za slučaj da je potrebno onemogućavanje ili uklanjanje dodatka s računara. Google Chrome povremeno preuzima ili ažurira ovu listu.

Ukoliko se koriste dodaci koje hostuje Google, pretraživač nam šalje jedan ili više jedinstvenih brojeva informacija kada se prvi put instalira dodatak, kada Google Chrome proveri da li ima ažuriranja za dodatak i kada ga deinstalirate. Ove informacije se koriste za praćenje korisničkih podataka o dodatku u galeriji dodataka i za rangiranje popularnosti dodataka.

7.10. GoogleBot

Googlebot je Google-ov inteligentni agent, koji sakuplja dokumenta sa Web-a i pravi adekvatan indeks za pretražovanje koji Google koristi. Googlebot koristi algoritamski proces: računarski programi određuju koje sajtove da pretražuju, koliko često i koliko stranica da donesu sa svake lokacije.

Ako Web-master želi da zabrani da informacije koje su na njegovom sajtu budu dostupne googlebot-u ili nekom drugom Web-pauku, to možu da uradi uz odgovarajuće direktive u „robots.txt“ fajlu ili dodavanjem meta taga „`<meta name="Googlebot" content="nofollow"/>`“ na Web stranicu. Googlebot zahtevi za Web servise su prepoznatljivi od strane korisnik-agent stringa koji sadrži „googlebot“ i host adresu koja sadrži googlebot.com.[27]

Googlebot je pratio HREF linkove i SRC linkove, ali od nedavno ima i sposobnost da analizira JavaScript, AJAX i Cascading Style Sheets. U prošlosti programeri su izbegavali korišćenje JavaScript-a da isporuči sadržaj ili linkove ka sadržaju zbog teškoće od strane Googlebot-a da pravilno indeksira ovaj dinamični sadržaj. Tokom godina Googlebot je postao toliko dobar u tom zadataku da Google sada traži da se omogući GoogleBot skeniranje JavaScript-a u sajtovima. Googlebot otkriva stranice tako što prikuplja sve linkove na svakoj strani koju pronađe . Zatim sledi ove linkove ka drugim stranicama . Nove Web stranice moraju biti povezane sa drugim poznatim stranicama na internetu kako bi se popisale i indeksirale ili bile ručno unete od strane Web-mastera .[28]

Problem koji Web masteri često imaju sa Googlebot-om je da zauzima ogromnu količinu protoka. Ovo može izazvati sajtu da pređe granicu svog propusnog opsega i da se skine privremeno. Ovo je posebno problematično za manje sajtove koji hostuju mnogo gigabajta podataka. Google nudi "Webmaster Tools" koji omogućava vlasnicima Web sajtova da smanje brzinu popisivanja.

[27] http://en.wikipedia.org/wiki/Googlebot
[28] http://www.beanstalk-inc.com/blog/2012/05/28/googlebot-now-indexing-javascript-ajax-css/

8. Zaključak

Najefikasniji vid pristupa sadržaja na Web-u, jeste korišćenje pretraživanja. Na ovaj način, izbegava se potreba za sistematskom organizacijom celokupnog Web-a, već se deo problema prenosi na samog korisnika, koji svojim upitima vrši selekciju segmenata kojima želi da pristupi. Zadatak Web pretraživača (web search engine) je da pruži mehanizme koji korisniku omogućavaju efikasno pretraživanje i pomoć pri selekciji relevantnog sadržaja. U poslednjoj deceniji, razvijen je značajan broj izuzetno efikasnih pretraživača, od kojih neki danas predtstavljaju najposećenjije lokacije na Web-u. Google karakteriše jednostavnost korišcenja, brzina odgovora i konstantan razvoj ponude, proizvoda i usluga. Navedena kombinacija pokazala se kao neverovatan korisnički magnet. Provođenjem spomenutog načina rada u stvarnost Google postaje sve popularniji među iskusnim, kao i neiskusnim; tj. novim korisnicima.

Optimizacija internet pretrage je dug proces koji postepeno poboljšava ranking u rezultatima pretrage, broj posetilaca i naravno vrednost (odnosno zaradu) nekog web sajta. Poboljšanje rankinga kod vodećih internet pretraživača se postiže sporo, po par meseci do preko godinu dana., tako da je za to vreme potrebno fokusirati se na lokalne pretraživače gde su mogući rezultati već posle nekoliko nedelja. Najjednostavnija provera relevantnosti pretrazivača jeste da otkucate ključne reči za koje znate najrelevantnije sajtove, bilo na engleskom ili nekom drugom jeziku. U zavisnosti od rezultata pretraživanja videćete i koliko su posmatrani pretraživači relevantni.

Internet pretraživači i pored pospešivanja, napredka i popularnosti WWW-a, doprinose i njegovoj kategorizaciji i organizaciji. Kad bi svi internet pretraživači prestali u trenutku sa radom, nastupio bih opšti haos na Internetu. Jednostavno rečeno, bez internet pretraživača ne bi bilo ni pretraživanja, već samo pukog traženja igle u plastu sena.

Danas, na početku 21. veka, zbog velike količine informacija koje nas okružuju, neophodno su nam potrebni inteligentni agenti koji omogućuju što bolje snalaženje tokom pretraživanja Web-a. Inteligentni agenti za pretraživanje javili su se u drugoj polovini osamdesetih godina, a njihovo izučavanje je intezivno počelo u devedesetim jer su prepoznati kao obećavajuće rešenje za problem prikupljanja, obrade i čuvanja velike količine informacija.

Očekivanja korisnika su mnogo narasla od doba kada su se pretraživački mehanizmi tek pojavili na Internetu i kada je bilo pravo čudo kada bi nešto pronašli. Danas, ako korisnik ne dobije što traži unutar prva tri linka, često odustaje od daljeg pretraživanja. Zato se svakodnevno, u velikim timovima, radi na ispravljanju algoritama pretraživanja i poboljšanju samog pretraživanja. Google-ov sistem za prikazivanje rezultata se godišnje nadogradi otprilike 500 puta.

Tolika upornost, svakodnevne izmene i dopune algoritama, veliki broj ljudi okupljenih u timu kako bi ispravljali greške koje se svakodnevno događaju i prikupljali pritužbe ljudi koji su razočarani rezultatima pretraživanja, ključ su za dobijanje tržišne utakmice. Naime, ljudi su postali razmaženi korisnici koji više ne traže odgovor na upit koji upišu, već zahtevaju da im se da ono što žele. Dakle, umesto "daj mi ono što sam upisao", sada je "daj mi ono što hoću".

Takođe je važno napomenuti da različiti problemi pretraživanja mogu biti rešeni primenom samo određenih algoritama pretraživanja, tako da ne postoji metoda pretraživanja koja može rešiti sve probleme pretraživanja. Stoga je tokom donošenja odluke koju metodu pretraživanja primeniti, važno proučiti samu kompleksnost problema.

Postoji i jaz između načina na koji ljudi šalju zahtev za informacijom i načina na koji računari pohranjuju informacije. Istraživanja na području obrade prirodnog jezika bi mogla znatno doprineti rešavanju ovog problema. Razumevanje jezika obuhvata razumevanje sadržaja i konteksta, a ne samo strukturu rečenice. Ukoliko bi ljudi mogli reći agentu šta žele prirodnim jezikom, a on to mogao da razume, pretraživanje bi imalo mnogo bolje rezultate.

Pretraživanje se može pomeriti sa sintaktičke osnove na konceptualnu ili semantičku osnovu dopuštajući agentu da traži ideje, a ne samo reči. Ukoliko agent razume koncepte (pojmove), onda se može koristiti učenjem kako bi mogao razumeti ponašanje korisnika i usavršiti pretraživanja za pojedini korisnički tip. Istraživanja na području planiranja, pretraživanja, integracije prirodnog jezika u računare i učenja računara mogla bi znatno doprineti poboljšanju inteligencije agenata za pretraživanja. Inteligentni agenti za pretraživanje zasigurno predstavljaju novi izazov u informatičkom svetu koji treba prihvatiti.

9. Literatura

[2] Tara Calishain (2007.): „Web Search Garage“

[6] John Battelle (2005.): „The Search“

[13] Nguyen N.T., L.C. Jain, (2009): „Intelligent Agents in the Evolution of Web and Applications“

[16]Germano Resconi, L. C. Jain, (2004): „Intelligent Agents: theory and applications“

Online izvori:

[1] http://en.wikipedia.org/wiki/Search_engine_technology

[3] http://en.wikipedia.org/wiki/Search_engine

[4] http://sr.wikipedia.org/sr/SEO_optimizacija_veb_sajta

[5] http://en.wikipedia.org/wiki/Search_engine_optimization

[7] http://www.cins.org.rs/?p=367

[8] http://members.tripod.com/~veliborb

[9] www.scribd.com/doc/39403486/inteligentni-agenti

[10] http://bs.scribd.com/doc/59442859/01-Vestacka-Inteligencija

[11] www.csc.liv.ac.uk/~mjw/pubs/ker95/ker95-html.html

[12] http://internet.fon.bg.ac.rs/download/piai/Agents%20-%20Slides.pdf

[14] www.acm.org/crossroads/xrds5-4/webhunting.html

[15] http://prelog.chem.pmf.hr/~tezak/preinin/preinin/index.html

[17] http://en.wikipedia.org/wiki/Intelligent_agent

[18] http://www.cs.berkeley.edu/~russell/aima1e/chapter02.pdf

[19] http://pespmc1.vub.ac.be/PROBSOLV.html

[20] www.acm.org/crossroads/xrds5-4/webhunting.html

[21] http://www.course.cs.york.ac.uk/lpa/slides2006/uninformed-search.pdf

[22] http://en.wikipedia.org/wiki/Knowledge_Query_and_Manipulation_Language

[23] http://en.wikipedia.org/wiki/Google

[24] http://radicalproject.blogspot.com/2011/06/googlati-nekoga.html

[25] http://www.zezas.me/2010/12/10-cinjenica-o-guglu-koje-verovatno.html?utm_source=linkeri.com&utm_medium=referral&utm_campaign=linkeri

[26] http://en.wikipedia.org/wiki/Google_Chrome

[27] http://en.wikipedia.org/wiki/Googlebot

[28] http://www.beanstalk-inc.com/blog/2012/05/28/googlebot-now-indexing-javascript-ajax-css/