Učinkovito pretraživanje Interneta
(alati i tehnike)

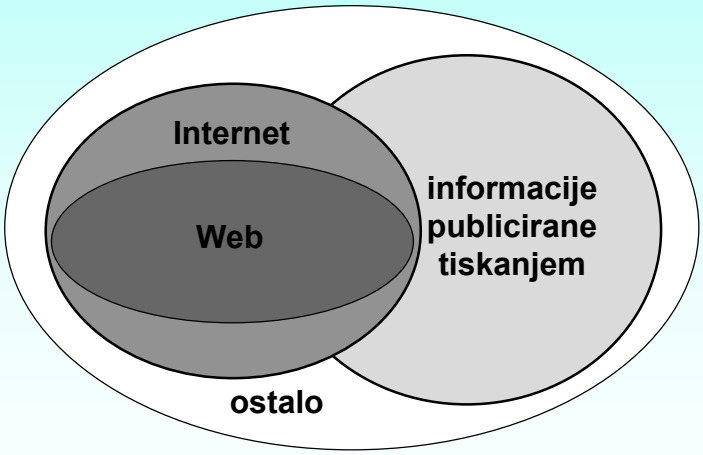
Miroslav Milinović
<miro@srce.hr>

Seminar za knjižnice visokih učilišta i znanstvene knjižnice, Zagreb, studeni, 2000.

Sadržaj

- Internetski prostor informacija
- Pretraživanje Weba
- Pretraživački mehanizmi (*Search Engines*)
- Tematski katalozi (*Subject Catalogs*)
- Ostali alati
- Portali
- Zaključak

Prostor informacija



Internetski prostor informacija

- NIJE UREĐEN - unificiran
- Postoje različiti izvori informacija (resursi)
- Mnoštvo tema
- Informacije su dostupne u različitim formatima
- Pristup je moguć pomoću različitih alata (programa)
- Postoje informacije koje (još) **nisu**:
  - publicirane u elektroničkom obliku
  - dostupne putem mreže

Web informacijski prostor

- pretraživi (*publicly indexable*) Web
  - veljača 1999., *Lawrence and Giles, NEC Institute*
    - 800 miliona stranica, 15 (6) TB informacija
    - sadržaj: 83% com, 6% sci/edu, 1.5% porn
    - 60% Weba je indeksirano / katalogizirano
  - siječanj 2000., *Inktomi & NEC Institute*
    - više od 1 milijarde Web stranica
    - top-level domene: 55% .com, 8% .net, 4% .org, 1% .gov

Web informacijski prostor

- 40% od 800 miliona stranica su duplikati

  FAST, 2000.
- 30% Web stanica su kopije

  Shivakumar and Garcia-Molina, 1998.
- "Deep" Web
  - 400 do 550 puta veći od "surface" Weba
  - 7500 TB podataka

    The Deep Web: Surfacing Hidden Value; BrightPlanet.com, srpanj 2000.

Web informacijski prostor

- 85% korisnika rabi pretraživačke mahanizme ili tematske kataloge kako bi pronašli informacije

  Steve Lawrence, Lee Giles , Nec Institute, veljača 1999.

- korisnici smatraju da je Internet važan izvor informacija
  - 2/3 korisnika smatra da je Internet važan ili vrlo važan izvor informacija
  - 53%(47%) smatra TV (radio) jednako važnim

    Center for Communication Policy, UCLA, kolovoz 2000.

Problemi?

- velika očekivanja korisnika
- alati i mehanizmi
  - još uvijek nedovoljno dobri
  - u stalnom razvoju
- informacijski prostor nije (dobro) organiziran
- nepouzdana:
  - kvaliteta informacija
  - integritet informacija
  - povjerenje u izvor informacija

Znate li ...

- **tko je bila prva žena pilot u nekoj komercijalnoj avio-kompaniji? Možete li pronaći njenu sliku (traži se točna URL adresa)?**

  •**Odgovor:** Helen Richey; da (http://iswap.org/images/richey.jpg)

  •**Put:** Rabimo **Northern Light** s upitom **"first woman airline pilot"**. Jedan od prvih 10 odgovora je i link na *ISAfaqs.html* Web stranicu.

  •**URL:** http://iswap.org/ISAfaqs.html

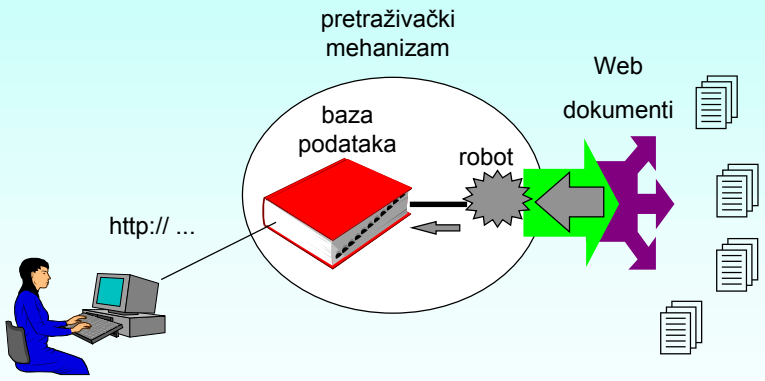
Alati za pretraživanje Weba

- **Pretraživački mehanizmi** (*search engines*)
  - pretraživački mehanizmi *(search engines)*
  - metapretraživački mehanizmi *(metasearch engines, unified search interfaces)*
- **Tematski katalozi** (*subject catalogs, subject indexes, subject directories, virtual libraries*, ...)
  - pretraživi (searchable indexes, searchable catalogs)
- **Ostali alati**:
  - višestruka sučelja *(multiple search interfaces)*
  - specijalizirana sučelja (*information gateways*)
  - **...**
- **Portali**

Pretraživački mehanizmi
Što su i kako rade?

- Automatizirani sustavi koji prikupljaju informacije o mrežnim resursima i omogućuju pretraživanje prikupljenih informacija
- Prikupljanje informacija obavljaju posebni programi - roboti (*robot, crawler, spider*)
  - robot pregledava dostupne mrežne resurse (Web dokumente)
  - gradi pretraživu kolekciju podataka (bazu podataka)
  - provjerava ažurnost izgrađene baze podataka i obnavlja njen sadržaj
- Web sučelje omogućuje korisniku pretraživanje baze podataka (*database search*, *index search*)

Pretraživački mehanizmi
Što su i kako rade? (2)



Pretraživački mehanizmi
Primjeri

GO.com (InfoSeek) - **http://www.go.com/**
Lycos Search - **http://www.lycos.com/**
Alta Vista - **http://www.altavista.com/**
excite! NetSearch - **http://www.excite.com/**
Google - **http://www.google.com/**
HotBot - **http://hotbot.lycos.com/**
WebCrawler - **http://www.webcrawler.com/**
Nothern Light Search - **http://www.northernlight.com/**
FAST - **http://www.alltheweb.com/**
Raging Search - **http://ragingsearch.altavista.com/**

pretraživački mehanizmi lokalnog dosega
http://cross.carnet.hr/

Pretraživački mehanizmi
Mogućnosti kod postavljanja upita

- uporaba malih i velikih slova
  `John December`
  `island`
- uporaba fraza
  `“John December”`
  `“NASA Space shuttle program”`
- uporaba logičkih operatora (AND, OR, NOT)
  `vegetables AND green`
  `fruit NOT apple`
- kontrola ključnih riječi (+, -)
  `+film +noir -”pinot noir”`
  `+python -monty`

☞

Pretraživački mehanizmi
Mogućnosti kod postavljanja upita (2)

- susjednost - proximity search
  `Internet NEAR training`
- uporaba dijelova (korijena) riječi (Keyword Truncation) - *, %
  alumi*um
  comput*
- kaskadno pretraživanje (Infoseek)
- kontrola resursa (AltaVista, HotBot, Infoseek)
  `title:"Internet training"`
- *natural language searching* (Ask Jeeves! - **http://www.ask.com/**)
- novi pristupi:
  - Ditto.com - **http://www.ditto.com/**
  - Simpli.com - **http://www.simpli.com/**
  - Oingo - **http://www.oingo.com/**

Pretraživački mehanizmi
Važne odlike

- Baza podataka (veličina, ažurnost, složenost)
  - Google - 560 million web pages
  - INKTOMI - 500 million web pages
  - AltaVista - 350 million web pages
  - FAST - 340 million web pages
- Mogućnosti postavljanja (složenih) upita
- Brzina rada (odziv)
- Rangiranje rezultata (*ranking*)
- Kvaliteta i mogućnost kontrole ispisa
- Dodatne mogućnosti
  (kaskadno pretraživanje, profinjavanje upita ...)

Pretraživački mehanizmi
Prednosti i mane

- Prednosti:
  - veliki opseg
  - efikasno pretraživanje i pristup informacijama
  - automatiziran rad
- Mane:
  - nema kontrole kvalitete
  - nema klasifikacije
  - rezultati mogu biti izvan konteksta (npr. “film”)
  - sadrže i zastarjele i nepostojeće URL adrese
  - sadrže i smeće

Pretraživački mehanizmi
Metapretraživački mehanizmi

- ***metasearch engines, unified search interfaces***
- omogućuju korisniku da putem unificirane forme postavi jedan upit kojeg zatim distribuiraju odabranim pretraživačkim mehanizmima
- kod postavljanja upita treba koristiti samo sintaksu koju poznaje metapretraživački mehanizam
- korisnik dobiva zbirni rezultat pretraživanja
- nemaju vlastite baze podataka niti robot program

Pretraživački mehanizmi
Metapretraživački mehanizmi (2)

- **primjeri metapretraživačkih mehanizama:**

  All4one - http://all4one.com/
  Mamma - http://www.mamma.com/
  MetaCrawler - http://www.metacrawler.com/
  SavvySearch (CNET Search.com) - http://www.savvysearch.com/

Pretraživački mehanizmi
Metapretraživački mehanizmi (3)

- **važne odlike**:
  broj i izbor povezanih pretraživačkih mehanizama
  brzina rada (odziv)
  rangiranje rezultata
  način udruživanja rezultata (*results merging*)
  kvaliteta ispisa
  mogućnost kontrole ispisa
  dodatne mogućnosti

Pretraživački mehanizmi
Metapretraživački mehanizmi (4)

- imaju sve prednosti i mane običnih pretraživačkih mehanizama
- **dodatna prednost:**
  - pojednostavljuju pristup i pretraživanje
- **dodatne mane:**
  - unificiranjem upita gube se dodatne mogućnosti postavljanja složenijih upita i kontrole ispisa
  - sporije pretraživanje

Tematski katalozi
Što su i kako rade?

- tematski organizirane kolekcije podataka o odabranim mrežnim resursima
  (odabrani resursi klasificirani po temama)
- sadrže URL adrese mrežnih resursa
- mogu sadržavati i nazive resursa, sažetke, ...
- ne održavaju se automatski (programski) već se temelje na radu urednika

Tematski katalozi
Što su i kako rade? (2)

- klasificiranje resursa se odvija prema hijerarhijskoj shemi tema (područja)
- način klasificiranja nije unificiran
  (UDC, Dewey, proizvoljan ...)
- postoji mogućnost pretraživanja kataloga

Tematski katalozi
Primjeri

Yahoo - **http://www.yahoo.com/**
LookSmart - **http://www.looksmart.com/**
EINet Galaxy - **http://galaxy.einet.net/**
Magellan - **http://magellan.excite.com/**
NetGuide - **http://www.netguide.com/**
About.com - **http://www.about.com/**
Open Directory - **http://dmoz.org/**
Brittanica.com - **http://www.brittanica.com/**

katalozi lokalnog opsega:
WWW.HR - **`http://www.hr/wwwhr/`**

Tematski katalozi
Važne odlike

- veličina (broj klasificiranih resursa)
  - Yahoo - 150 urednika, 1.2 miliona Webova (1999.)
  - Open Directory - 31,095 urednika, 2,1 miliona Webova (2000.)
- tematsko stablo - način klasifikacije
- dostupne informacije o resursima
- rangiranje resursa
- mogućnost pretraživanja
- dodatne mogućnosti
- ...

Tematski katalozi
Prednosti i mane

- Prednosti:
  - klasifikacija resursa po temama (područjima)
  - mogućnost internog pretraživanja kataloga
  - nema “smeća”
- Mane:
  - manualno održavanje
  - pojedine dijelove kataloga ne uređuju profesionalci
  - sadrže i zastarjele informacije

Ostali alati

Višestruka sučelja (*multiple search interfaces*)

- jednostavna sučelja koja korisniku omogućuje da na jednom mjestu odabere pretraživački mehanizam
- nemaju vlastite baze podataka niti robot program
- primjeri:
  - All-in-One - **`http://www.albany.net/allinone/`**
  - Easy Searcher - **`http://www.easysearcher.com/`**

Ostali alati (2)

Specijalizirana sučelja (*information gateways*)

- **prednosti:**
  - korektno klasificiran sadržaj uvijek u kontekstu
  - moguće pretraživanje
- **mane:**
  - vezani uz jednu temu (područje)
  - manualno održavanje
- **primjeri:**
  OMNI - **`http://www.omni.ac.uk/`**
  SOSIG - **`http://sosig.ac.uk/`**

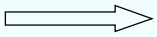
Ostali alati

- **Imenički servisi utemeljeni na Webu**
  - White pages & Yellow pages
  - ne rabe niti LDAP niti neki drugi protkol specifičan za imeničke servise
- **Web alati za pretraživanje ne-Web resursa**
  - USENET (http://www.deja.com/usenet/)
  - FTP search (http://ftpsearch.lycos.com/)
  - distribucijske (mailing) liste (http://www.liszt.com)
  - . . .

Ostali alati (4)

- **pretraživanje kolekcija (baza) podataka**

  Inivisible Web - **http://www.invisibleweb.com/**
  Lycos Seach. DB - **http://dir.lycos.com/Reference/Searchable_Databases/**
  INFOMINE - **http://infomine.ucr.edu/**
  Terraserver - **http://terraserver.com/**
- **i ...**
  - rječnici, enciklopedije, vodiči, pretražive kolekcije multimedijalnih sadržaja, ….



PORTALI

Portali

- ulaz u informacijski prostor Interneta
- hibridni alat - pravo rješenje
- nude pristup (svim) mrežnim servisima na jednom mjestu
- temelje se na pretraživačkom mehanizmu i/ili tematskom katalogu
- opći ili specijalizirani (tema ili interesna skupina)
  - **http://cnn.com/**
  - **http://www.excite.com/**
  - **http://www.altavista.com/**
  - **http://www.yahoo.com/**
  - **http://www.ihlth.com/**
  - **http://www.digitalessays.com/**
  - ...

Alati za pretraživanje Weba
Zaključak

- svaka grupa alata ima svojih prednosti i mana
- orijentirani su na tekst dokumenta
  (slikovni i zvučni zapis nije moguće pretraživati po sadržaju)
- očekuje se da obuhvaćaju i ne-Web resurse
- temeljne brige:
  - kako biti ažuran
  - kako očuvati kvalitetu (precision .vs. recall)
  - kako odijeliti "mrežno smeće" od kvalitetne informacije
- budućnost je u "suradnji među alatima"
- pobjednik: **PORTAL**
- korisna adresa: http://searchenginewatch.com/

Pretraživanje Web resursa
Izbor alata

- **PORTALI !**
- **tematski katalozi**
  - kad nemamo (dobre) ključne riječi odnosno jasnu ideju što tražimo
- **pretraživački mehanizmi**
  - kad imamo precizne ključne riječi i jasnu ideju što tražimo
- **višestruka sučelja**
  - korisna jer daju pregled raspoloživih alata
- **specijalizirana sučelja (za neko područje)**
  - nude kvalitetne informacije (ako postoje i znamo za njih)

Pretraživanje Web resursa
Kako pretraživati?

- dobar izbor ključnih riječi je presudan
- biti usmjeren k cilju (Ne lutati!)
- treba se koncentrirati na temu, a ne na postavljanje uputa
- ići k cilju postepeno (profinjavati upite)
- upoznati alat (Pročitajte HELP i FAQ!)
- biti fleksibilan i probati više različitih (tipova) alata
- graditi vlastite kolekcije zanimljivih mjesta na mreži

O čemu je bilo riječi?

- Internetski prostor informacija
- Pretraživanje Weba
- Pretraživački mehanizmi (*Search Engines*)
- Tematski katalozi (*Subject Catalogs*)
- Ostali alati
- Portali
- Zaključak