Preddiplomski studij

Računarstvo

Modul:
Telekomunikacije i informatika

Višemedijske usluge

Pretraživanje informacija na WWW-u (2. dio)

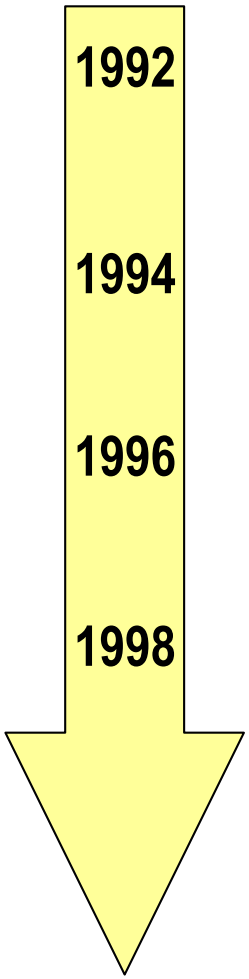
Ak.god. 2007./2008.

24.04.2008.

Sadržaj predavanja

- Razlike u odnosu na “klasični IR”
- Arhitektura Web tražilice
- Rangiranje (PageRank)
- Višemedijske tražilice

Malo povijesti...



Počeci weba

preglednici

Imenici

- Yahoo

Prve tražilice

- InfoSeek, Lycos, Altavista, Excite, Inktomi, …

Preporod web tražilica

- Google

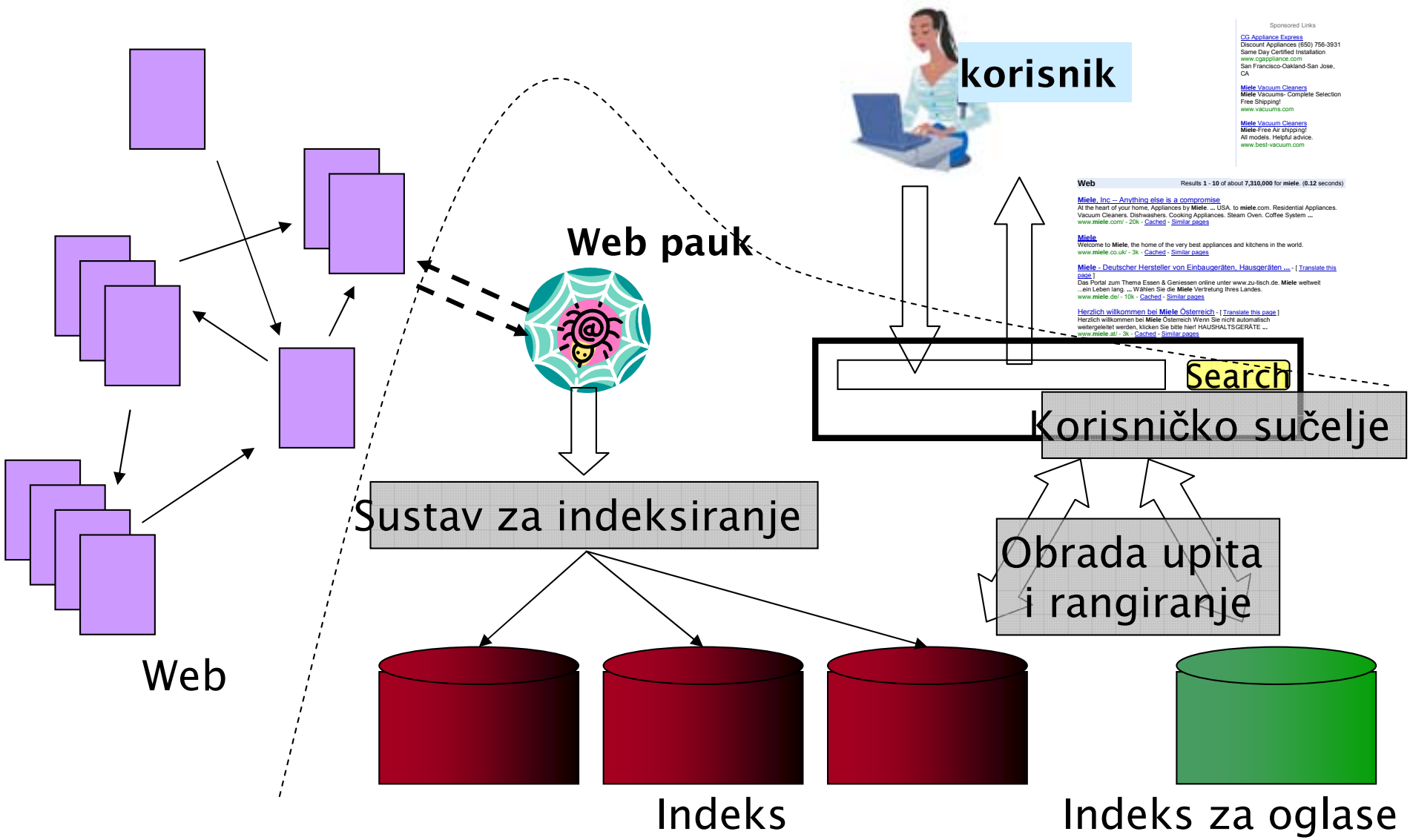
Razlike u odnosu na “klasični IR”

- Kolekcija
  - veličina, dinamične promjene dokumenata
  - velike razlike u kvaliteti dokumenata
  - velika količina “duplikata”
  - velika količina sadržaja na Webu nije indeksirana (*deep Web*)
- Korisnici
  - postavljaju kratke upite (najčešće 2 do 3 riječi)
  - neprecizno definirane potrebe za informacijama

Zahtjevi

- jednostavno korisničko sučelje
- kratko vrijeme odziva
- rangiranje rezultata je iznimno važno
  - većina korisnika ne koristi rezultate nakon prve stranice
- važna je preciznost rezultata na prvoj stranici odgovora
- odziv je vrlo teško ocijeniti

Arhitektura Web tražilice



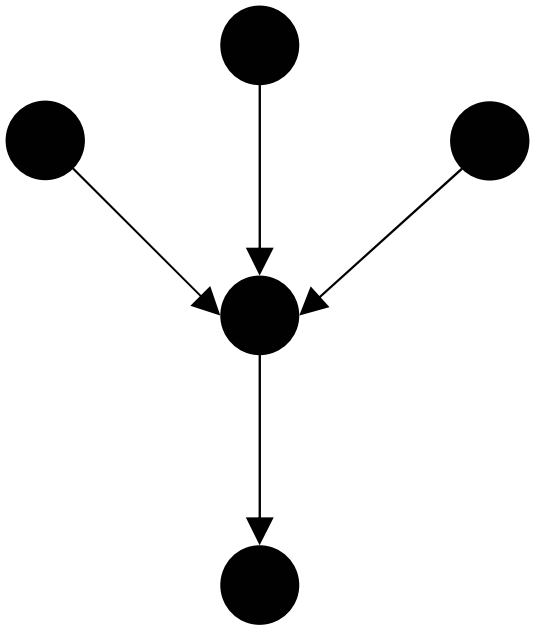
Dijelovi tražilice

- Pauk (engl. *spider*, *crawler*)
  - program koji stvara kolekciju tako što posjećuje poznate web stranice, analizira sadržaj stranice te prati ugrađene poveznice
  - slika Weba koju danas pretražujemo poznatim tražilicama stara je oko mjesec dana
  - Što je s vijestima, blogovima?
- Sustav za indeksiranje
  - kreira distribuirani invertirani indeks
- Sustav za obradu upita i rangiranje
  - implementira model za pretraživanje
  - česte riječi se ignoriraju, a ostale svode na korijenski oblik (stemming)

Rangiranje

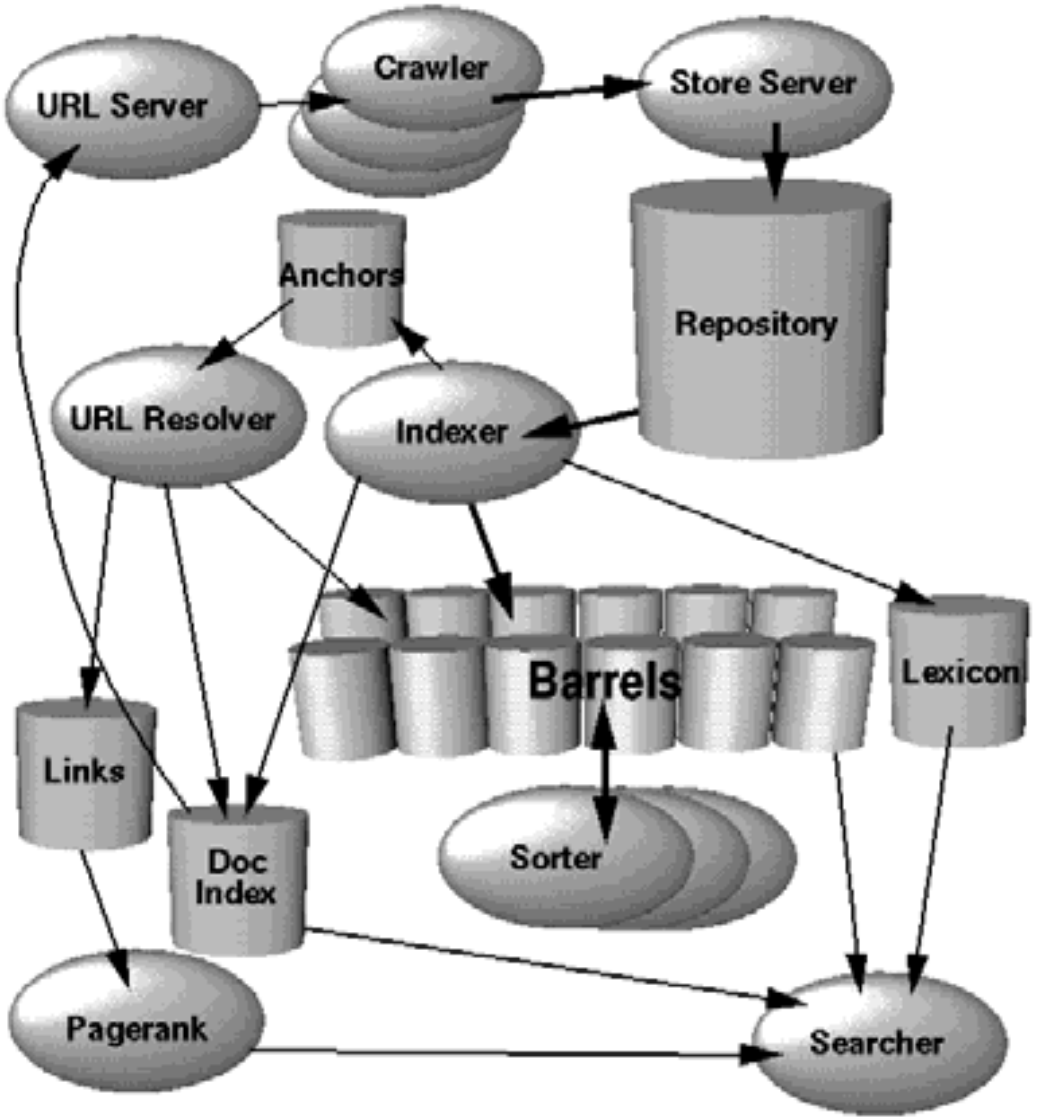
PageRank

- algoritam koji je učinio Google najpopularnijom tražilicom
- modelira Web usmjerenim grafom
- koristi ulazne i izlazne poveznice radi rangiranja relevantnih stranica s obzirom na njihovu popularnost
- neovisan o upitu



Ako vektorski model rangira 2 stranice jednako, PageRank će dati prioritet popularnijoj stranici, tj. stranica s više ulaznih poveznica dobiva viši PageRank pogotovo ako te stranice imaju veliki PageRank.

Google



© Sergey Brin and Lawrence Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine, 1998.

Najpoznatije tražilice

- Google, http://www.google.com/
- Yahoo!, http://search.yahoo.com/
- Ask.com, http://www.ask.com/
- Altavista, http://www.altavista.com/
- Pogodak!, http://www.pogodak.hr/

Popularnost tražilica

U.S. Core Searches by Search Engine, December 2007

Core Search Entity	November 2007 (%)	December 2007 (%)	Point Change
Total core search	100.0	100.0	0.0
Google sites	58.6	58.4	-0.2
Yahoo sites	22.4	22.9	0.5
Microsoft sites	9.8	9.8	0.0
Time Warner network	4.5	4.6	0.1
Ask network	4.6	4.3	-0.3

Izvor: Enid Burns, Search Engine Watch, Feb 5, 2008.
http://searchenginewatch.com/

Definiranje upita (1)

- 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

Definiranje upita (2)

- susjednost - proximity search
  `Internet NEAR training`
- uporaba dijelova (korijena) riječi (Keyword Truncation) - *, ?, %
  `alumi*um`
  `comput*`
- kontrola resursa
  `title:"Internet training"` (AltaVista, HotBot, …)
  `host:www.fer.hr` (AltaVista)
  `image:slika.jpg` (AltaVista)
  `related:` (Google)

Problemi postojećih tražilica

- najpoznatije tražilice poslužuju milijun upita dnevno a pretražuju više milijardi dokumenata (mali postatak cijelog Weba)
- premda sadržaj na Webu nastaje potpuno raspodijeljeno, pretraživanje toga sadržaja još je uvijek centralizirano
- problemi vezani uz generirani promet, sadržaji koji se često mijenjaju, sadržaj zaštićen autorskim pravima

Ostali sustavi za pretraživanje Weba (1)

- **metatražilice**
  - istovremeno pretražuju nekoliko tražilica, ali ne uključuju sve rezultate s pojedinih tražilica
  - nude jednostavne tehnike pretraživanja
  - korisne su za uspoređivanje različitih tražilica i dobivanje informacija postoji li na webu išta o traženoj temi

Ostali sustavi za pretraživanje Weba (1)

- **imenici** (*subject catalogs, subject directories, ...*)
  - baza podataka mrežnih stranica koje odabiru stručnjaci i organiziraju prema kategorijama hijerarhijski
  - pogodni su za općenita pretraživanja kada korisnik nema jasnu ideju što traži
  - u pravilu pretraživi (searchable indexes, searchable catalogs)
- Najpoznatiji imenici:
  - Google Directory, http://www.google.com/dirhp
  - Yahoo!, http://dir.yahoo.com
  - About.com, http://www.about.com
  - Infomine, http://infomine.ucr.edu
  - Academic Info, http://www.academicinfo.net

Višemedijske tražilice

Mutimedia IR

- tražilice specijalizirane za višemedijske dokumente (slike, glazba...)

- pretraživanje po sadržaju trenutno u eksperimentalnoj fazi
  - primjeri upita
  - Treba mi slika londonske telefonske govornice
  - Zanima me drugi gol s utakmice ...
  - Želim pronaći dio filma kada ...

Pretraživanje slika pomoću teksta (1)

Sign in

Google

Web **Images** Video News Maps more »

sunset Search Images Search the Web Advanced Image Search Preferences

Moderate SafeSearch is on

Images Showing: All image sizes

Results **1** - **20** of about **5,550,000** for **sunset** [definition]. (**0.04** seconds)

Sunset. **Sunset**
700 x 525 - 103k - jpg
www.whatdigitalcamera.com

...
/images/**Sunset**%2027.08.2005.jpg
1417 x 942 - 140k - jpg
home.online.no

sunset. ...
3070 x 2044 - 485k - jpg
www.solarviews.com

sunset.jpg 234123 bytes
800 x 600 - 229k - jpg
oz.irtc.org

The End of Another Day, **Sunset** Over ...
500 x 375 - 42k - jpg
seniortravel.about.com

Seattle **Sunset** background image
1024 x 768 - 110k - jpg
www.zenhaiku.com

Sunset Beach
448 x 336 - 21k - jpg
www.adrhi.com

beach **sunset**
500 x 375 - 66k - jpg
www.pixelcharmer.com

Imagine a **sunset**.
600 x 450 - 171k - jpg
www.grographics.com

Sunset at Rocky Point
442 x 302 - 31k
www.theodora.com

Pretraživanje slika pomoću teksta (2)

- najčešće koriste tradicionalne tekstualne metode za pretraživanje opisa vezanih uz slike
- ručno ili automatsko obilježavanje slika
- primjeri tražilica
  - http://picsearch.com
  - http://www.pixsy.com

Pretraživanje slika po sadržaju

- uspoređuje sličnost 2 slike po
  - boji
  - teksturi
  - obliku
- upit se definira kao slika-predložak
  - http://tiltomo.com
  - http://labs.systemone.at/retrievr/
- upit kao skica
  - http://hermitagemuseum.org/fcgi-bin/db2www/qbicSearch.mac/qbic?selLang=English

Pretraživanje videa

- Temelji se na tekstualnom pretraživanju metatagova
  - title, date, actors, produces, genre, running time, reviews, ratings...
- Primjeri tražilica
  - http://www.archive.org/
  - http://www.open-video.org/