don't be evil

Dobrica Pavlinušić
dpavlin@rot13.org
http://www.rot13.org/~dpavlin/

Kako u nestrukturiranom svijetu naći nestrukturirane informacije?

Don't be evil?

- Sjećate li se Altaviste?
- Sjećate li se kako je u izgledao Yahoo?
- .com boom je sve promjenio
- Internet više nije samo za power usere
- previše informacija
- kako naći informaciju koju trebam?
- semantic web je jedan od načina...
- Google je drugi...

Ali, sve je počelo sa...

informacijom

(ne)strukturirane informacije

- strukturirane informacije
  - baze podataka (u širem smislu)
  - riječnici i kazala
  - bibliografski podaci
  - telefonski imenici
  - oglasi
- nestrukturirane informacije
  - dokumenti, slike, zvuk, filmovi
  - moj stol (u neredu)

karakteristike

- strukturirane informacije
  - točno razdjeljuju pojmove i fraze
  - međusobne veze između pojmova
  - uređenje po nekoj logici ili abecedi
  - fiksan broj polja sa podacima
  - ponekad kontrolirani rječnik za pojmove
- nestrukturirane informacije
  - jednostavniji i bogatiji za izražavanje
  - višestruko veći broj korisnika

pregledavanje

- strukturirane informacije
  - prema određenoj strukturi
  - prema međusobnim vezama
  - prema vremenskim razdobljima
  - prema skupu podataka (filtriranje i agregiranje)
  - kombinacijom
- nestrukturirane informacije
  - prema eventualnim vezama (web)
  - jedan po jedan

pretraživanje

- strukturirane informacije
  - relativno mala količina
  - točan oblik pojma
  - ordeđene ključne riječi
  - logički operatori
  - jezik za pretraživanje
- nestrukturirane informacije
  - ogromna količina podataka (web i više)
  - nekoliko riječi iz punog teksta

Semantički web

- točno opisan odnos između različitih pojmova i stranica na webu
- Tim Berners-Lee (HTML, HTTP)
- zahtjeva od korisnika definiranje veza
- jednostavno i veoma moćno pretraživanje

Da li je lakše natjerati ljude da opisuju veze ili naučiti računalo da samo shvati veze?

Internet je pun...

nestrukturiranih informacija

Kako pretražiti Internet?

- instalirate nekoliko mašina
- napišete software koji skida sve stranice
- napravite indeks svih riječi i pretražujete ga
- napravite formu u koju korisnici upisuju upit
- prikažete korisnicima rezultate

Po čemu je Google drugačiji?

Želim pretražiti Internet!

- korisnik
  - forma sa poljem za upis upita
  - rangirani rezultati
- Google
  - računala
  - kopija cijelog sadržaja weba
  - indeks svih riječi
  - pretraživanje
  - povezivanje sa strukturiranim podacima

PageRank

- algoritam iza rangiranja rezultata
- poredak po broju citiranja stranice (broju linkova koji na webu vode prema stranici)
- više povezane stranice su ”važnije”
- nepovezani ili najnoviji dijelovi weba ostaju nepretraženi
- mogući pokušaji prijevare (for fun and profit)

Pretraživanje

- prema nekoliko riječi (1-3 u prosjeku)
- operatori (AND, OR)
- sa ili bez riječi (+knjižnice, -google)
- prema frazi (”pretraživanje interneta”)
- prema URL adresi stranice (site:hr)
- linkovima prema stranici (link:www.szi.hr)
- Advanced search sučelje
- Spremljena (cache) verzija stranice

Sitnice u pretraživanju

- similar pages
- kako se ono piše? (Did you mean?)
  - nije riječnik, samo statistički uzorak!
- ograničavanje po jeziku (Language Tools)
- kalkulator i konvertor
  - 30 + 12, 1 ft, 3 tea spoons, 10 in + 25 cm
- strukturirani podaci
  - US adrese, dionice, UPS, FedEx, FCC...

news.google.com

- mnogo brži ciklus obnavnjanja nego za web sadržaj
- izbor iz sredstava ”javnog informiranja”
- nema hrvatskih sadržaja
- Google News Alert – dobivajte obavijesti e-mailom

groups.google.com

- Usenet – mrežne vijesti
- hijerarhijska podjela sadržaja po temama
  - npr. hr.org.ffzg, hr.comp.linux
- veza između pojedinih postova
- pojedini rezultati mnogo specifičniji nego na webu (točniji!)
- tamna povijest Useneta
  - www.archive.org za stare web stranice

images.google.com

- pretraživanje slika
- ali, računalo ne ”vidi” sliku!
- pretraživanje po nazivu datoteke, okolnim pojmovima...
- filter za ”nepoćudan” sadržaj
- Picasa – pretražuje slike na vašem disku

gmail.google.com

- ”još da mogu tako pretražiti i svoj e-mail”
- webmail sa pretraživanjem i konceptom razgovora
- automatsko označavanje maila
- 1Gb prostora – nikada ne brišite e-mail!
- članstvo moguće samo uz poziv (ali imamo ih nekoliko)
- pitanja privatnosti

print.google.com

- dio testnog, beta programa
- upit: books on digital photography
- pretraživanje punog teksta knjiga
- bez mogućnosti ispisa ili kopiranja testa
- veoma slično Amazonom "search inside this book"
- nije DRM, ali je izdavačima dovoljno sigurno za većinu korisnika

desktop.google.com

- ”jednostavnije mi je nešto naći na Internetu nego ma mom računalu”
- pretražite svoje računalo!
- samo 500K (0.5Mb) datoteka
- pretražuje Outlook, presurfani Web, Word, PowerPoint, Excel i tekstualne datoteke na disku
- integracija sa internet searchom (uz usporenje!)

adwords.google.com

- od nečega se mora i živjeti...
- oglasi prema ključnim riječima

More informacija

- pronađite ono što tražite
- snađite se u njemu
- pomognite svojim korisnicima da se snađu

Nemojte se navući na pretraživanje kao ja!

Pitanja? I nadam se odgovori...