Veb *crawling* i analiza linkova

Dragan Ivanović
dragan.ivanovic@uns.ac.rs

Katedra za informatiku, Fakultet tehničkih nauka, Novi Sad

2015.

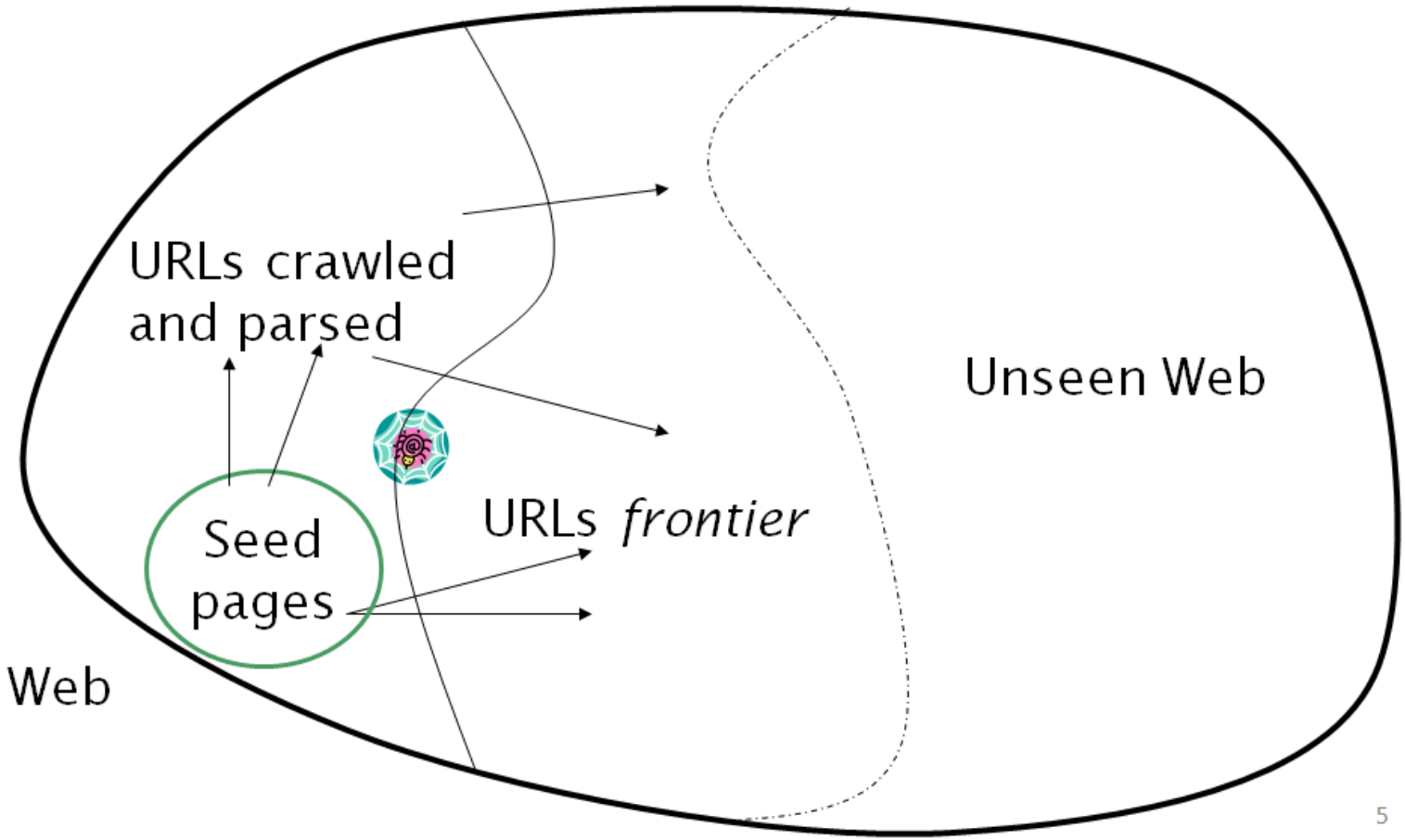
Osnove

- *Crawler* (poznat i kao *Spider*, odnosno *Web robot*) je program za automatsko kretanje kroz graf veba i preuzimanje veb stranica radi neke dalje obrade
- Obično se ta dalja obrada odnosi na indeksiranje zarad pretraživanja
- Graf veba je ogroman što ovaj zadatak prilično komplikuje
- Stranice na vebu se svakodnevno menjaju, što još više komplikuje zadatak
- Postoje i dinamički generisane stranice

Osnovne operacije

- *Crawler* počinje sa poznatim "*seed*" *URL*-ovima
- Preuzme ih i parsira
- Ekstrahuje *URL*-ove iz ovih fajlova
- Postavi ekstrahovane *URL*-ove u red
- Preuzme svaki URL iz reda i ponavlja prethodne stavke

Osnovne operacije



Komplikacije

- Nemoguće sa jednom mašinom - sve operacije su distribuirane na više mašina
- Maliciozne veb strane - spam strane, *spider traps*
- I sa nemalicioznim veb stranama postoje problemi - protok ka i kašnjenje odgovora sa udaljenih servera varira, koliko duboko u hijerarhiju sajta treba ići, *site mirrors* i približni duplikati
- Politeness (učtivost) - ne gađaj server zahtevima previše često

Šta se mora

- *Crawler* mora biti **učtiv** - poštovati implicitne i eksplicitne zahteve
  - Samo *crawl allowed pages*
  - Poštovati robots.txt
- *Crawler* mora biti **robustan** - mora na odgovarajući način da reaguje na nenormalne situacije, maliciozne strane, klopke za *Crawler*-e

Učtivost

- Eksplicitna - specifikacija sta može biti preuzeto *crawler*-om u propisanoj formi (robots.txt)
- Implicitna - čak i da nema specifikacije, mora izbegaviti preuzimanje puno strana u kratkom vremenskom roku

robots.txt

- Protokol za ograničenja pristupa crawler-a ("robota") veb sajtu - nastao 1994 godine (protokol)
- Veb sajt ovde navodi šta crawler sme, a šta ne sme
- Fajl robots.txt se stavlja u root veb sajta

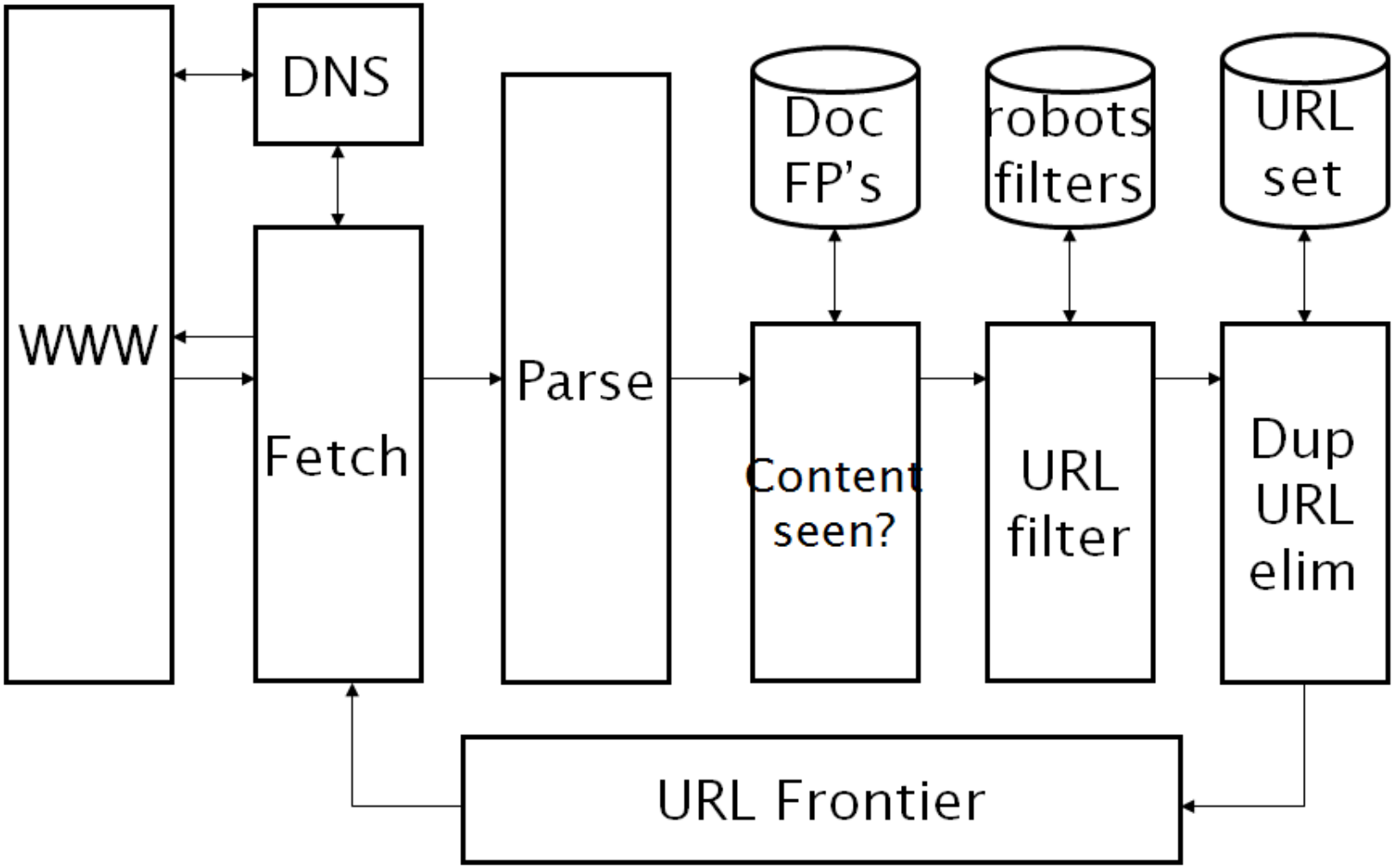
```
User-agent: *
Disallow: /yoursite/temp/

User-agent: Google
Disallow:
```

Šta bi bilo dobro

- **Distribuiranost** - da postoji mogućnost distribuiranog izvršavanja operacija, odnosno *crawler* dizajniran da se izvršava na više računara
- **Skalabilnost** - da se može dodavati još računara u sistem ako za to bude potrebe
- **Performanse (efikasnost) makimalne** - najbolja moguća iskorišćenost procesnih i mrežnih resursa
- **Prioritetnije preuzimanje** strana dobrog kvaliteta
- **Ponovno preuzimanje** novih sadržja sa strana koje su u međuvremenu izmenjene
- **Proširivost novim formatima i protokolima**

Osnovna arhitektura



URL frontier

- Sadrži *URL*-ove koje je potrebno preuzimati
- Može uključiti više strana sa istog *host*-a
- Mora voditi računa da se izbegava njihovo preuzimanje u istom trenutku (učtivost) - definiše se minimalni *time gap* između dva zahteva istom *host*-u
- Mora pokušati da u svakom momentu koristi sve resurse *crawler*-a (Performanse, efikasnost makimalne)

Naredni URL za preuzimanje

- Preuzmi *URL* iz *URL frontier*-a - koji URL?
- Preuzmi dokument sa tog *URL*-a, proveri da li se već nalazi u indeksima, ako ne onda ga indeksiraj
- Parsiraj dobijeni dokument i ekstrahuj *URL*-ove ka drugim dokumentima
- Za svaki ekstrahovani *URL* proveri da li prolazi sve filtere (samog *crawler*-a i veb sajta - robots.txt) i da li je već u *URL frontier*-u

URL normalizacija

- Kada se preuzeti dokument parsira neki od linkova su relativni
- Relativni linkovi se prebacuju apsolutne linkove pre nego što se stave u *URL frontier*

DNS

- Domain name server
  - Za simboličko ime vraća IP adresu
  - Distribuiran sistem - moguć spor odziv, nekoliko sekundi
- DNS *lookup*-i su blokirajući
- DNS keširanje - ako se jedanput traži IP adresa za neko simboličko ime rezultat se kešira
- *Batch DNS resolver* – prikupljanje zehteva i njihovo grupno slanje

Sadržaj već viđen

- Duplikata i približnih duplikata ima puno na vebu
- Ako je sadržaj preuzete strane već indeksirane dalje se ne procesira
- Provera se vrši poznatim tehnikama za utvrđivanje duplikata: *fingerprinting algorithms* ili *shingles* za približne duplikate

URL filteri

- Regularni izrazi kojima se definiše koje *URL*-ove *crawler* treba da preuzme radi indeksiranja - na primer želimo samo domen .edu
- Pored filtera definisanih u samom *crawler*-u, mora se poštovati i *robots.txt* preuzet za određeni veb sajt
- Voditi računa da se *robots.txt* preuzima samo jednom, keširati ga, ponovno preuzimanje je još jedan zahtev za veb server (setite se implicitne učtivosti)

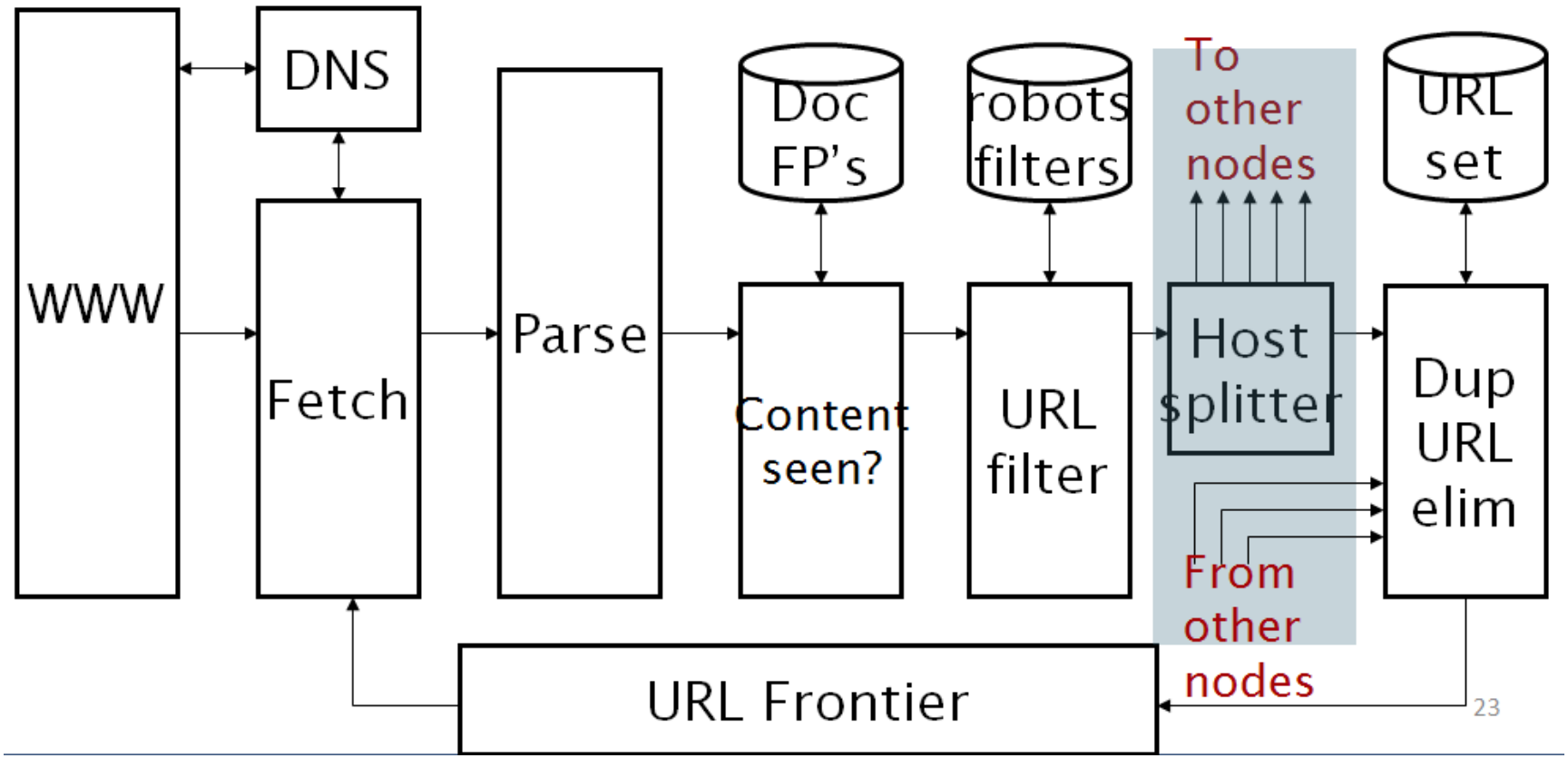
Eliminacija duplih *URL*-ova

- Kada crawler treba da jedanput preuzme sadržaje onda je ova komponenta vrlo prosta
- Ako je potrebno da periodično preuzima sadržaje kako bi imao njihove ažurne verzije (slučaj kod veb pretraživača) onda je to nešto složenije
- *Freshness* - preuzimaj neke strane češće od drugih (na primer sajtove sa dnevnim vestima koje se često menjaju)
- U ovom slučaju *URL frontier* ima složeniju arhitekturu
  - Poznata je *Mercator URL frontier scheme* - link
  - *front queues* - vode računa o prioritetima (*freshness*)
  - *back queues* - vode računa o učtivosti

Distribuiranost operacija

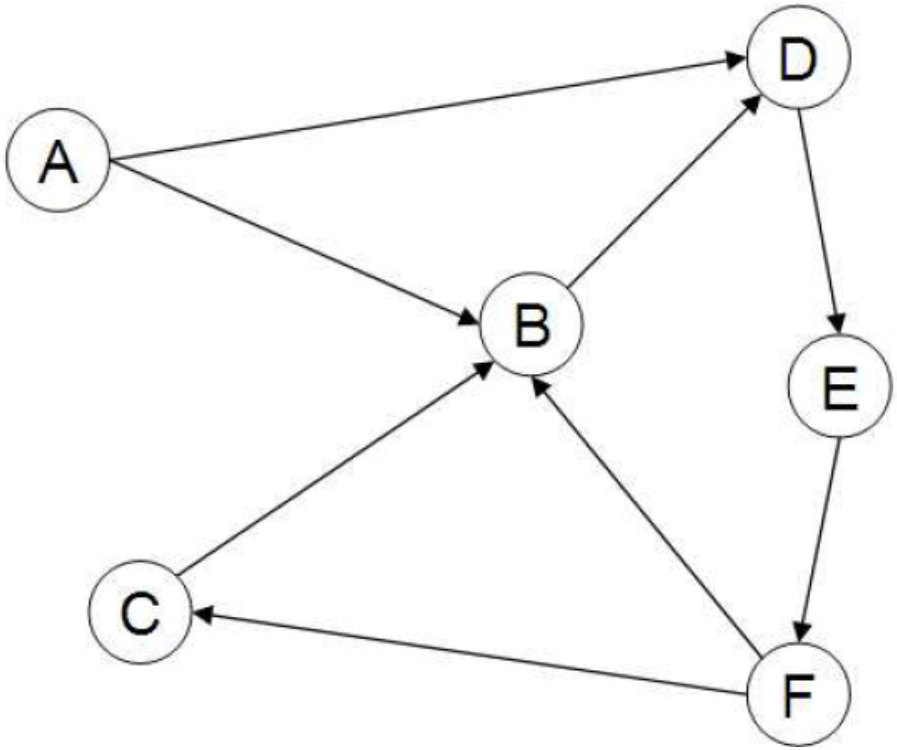
- Cilj - izboriti se sa ogromnim grafom veba
- Pokrenuti više niti na različitim mašinama (*node*-ovima) koje mogu biti i fizički udaljene
- Kako komuniciraju ovi node-ovi i ko je zadužen za koji URL?

Distribuirana arhitektura



Graf veba

- Na vebu pored sadržaja dokumenata imamo i linkove kojima su dokumenti povezani, moguće je formirati graf čiji su čvorovi dokumenti na vebu, a relacije su linkovi između njih, onda se može koristiti teorija grafova za neke zaključke



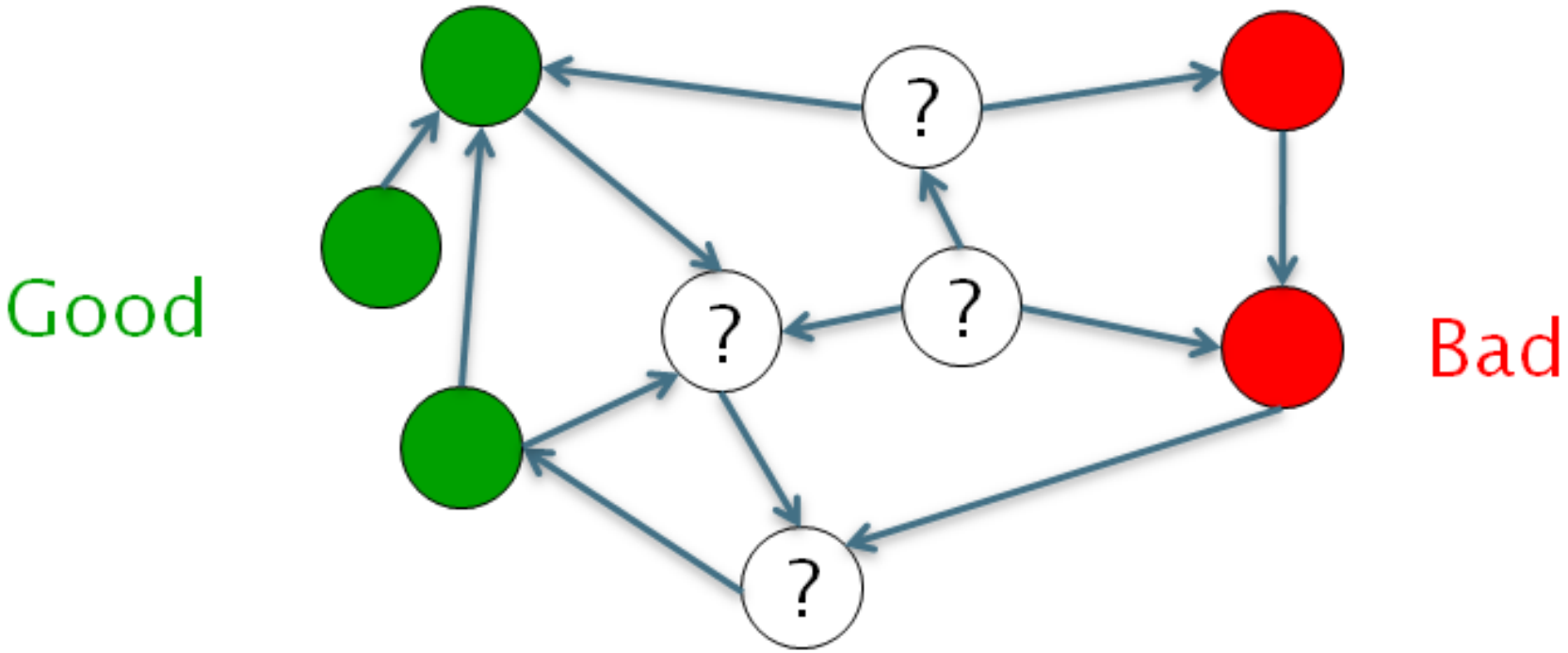
Linkovi

- Da li količina linkova ka nekom dokumentu ukazuje na značaj i kvalitet tog dokumenta ili te veb strane?
- Da li se linkovi mogu uključiti u rangiranje rezultata?
- *How likely is it that a page pointed to by the CERN home page is about high energy physics*

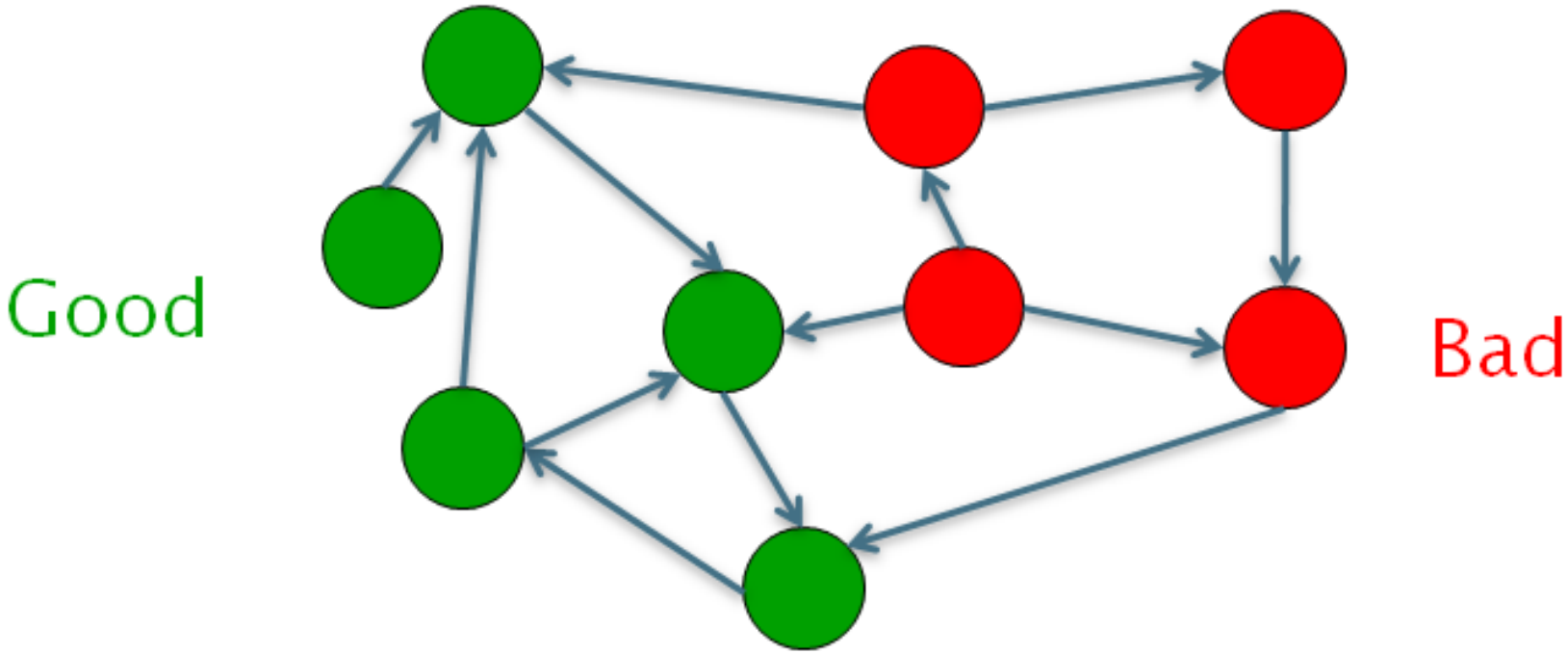
Dobri i loši čvorovi

- Podsećanje: korisnici žele pouzdane odgovore
- U veb grafu postoje dobri i loši čvorovi, ali i čvorovi za koje se ne zna da li su dobri ili loši
- Dobri čvorovi nemaju linkove ka lošim čvorovima, a sve ostale kombinacije su moguće
- Ako čvor ima linkove ka lošim čvorovima i on je loš
- Ako dobri čvorovi imaju link ka nekom čvoru i taj čvor je dobar
- U praksi je to mnogo složenije

Dobri i loši čvorovi



Dobri i loši čvorovi

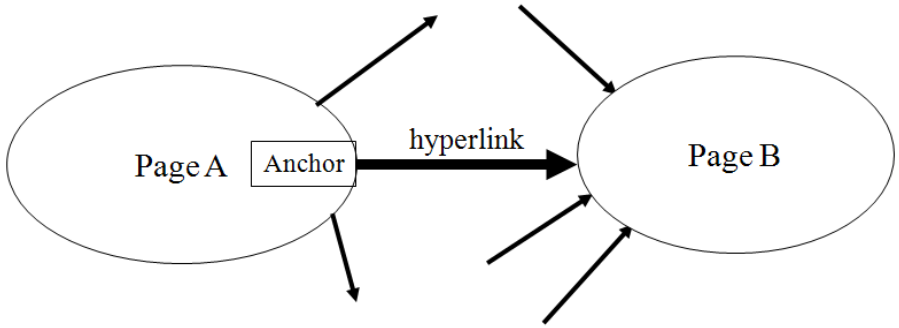


Primena link analize

- Utvrđivanje kvaliteta veb strane, odnosno primena u rangiranju rezultata kod pretrage veba, klasterovanje veb strana, klasifikacija veb strana, *crawling*
- Društvene mreže - grupe ljudi sličnih interesovanja
- Nauka - veze između radova
- Pronalaženje potencijalnih kupaca - onaj ko ima puno prijatelja koji troše mnogo, i sam troši mnogo
- Interesantna knjiga - link

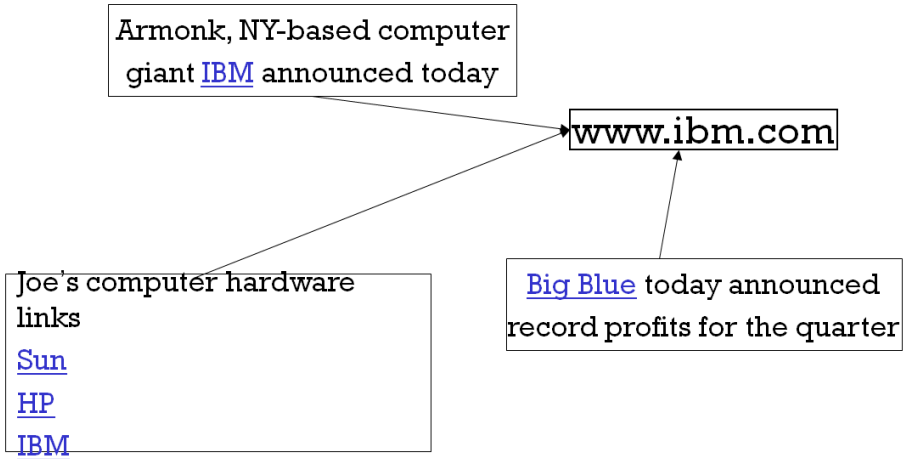
Usmereni graf veba

- Hipoteza 1. Link ka nekoj strani doprinosi meri kvaliteta strane
- Hipoteza 2. Tekst na stranici A nad kojim je definisan link (*anchor text* - tekst linka) opisuje stranicu B



Tekst linka

- Kada se indeksira dokument (ili veb strana) B, uključi (sa određenom težinom) tekstove linkova na stranicama koje upućuju na B
- Dokument B može biti indeksiran, može se i naći među odgovorima čak i ako dokument B nije (još uvek) preuzet *crawler*-om



Indeksiranje teksta linka

- Nekad je jasna fraza link na drugi dokument, a nekad je potrebno uzeti rečenicu ili deo rečenice kojoj link pripada.
- Nekad ima neočekivane bočne efekte - **evil empire** (Microsoft)
- Potrebno je dati težinu tom tekstu shodno kvalitetu veb stranice A sa koje link upućuje na stranicu B - teksta linka sa cnn.com je dobar za indeksiranje te stranice na koju se upućuje, tekstovi relativnih linkova imaju manju težinu

Druge primene teksta linkova

- Davanje težine vezi u veb grafu na osnovu teksta
- Opis strane na osnovu teksta linkova koji upućuju na tu stranu
- Klasifikacija veb strana, znamo klasifikaciju strane koja upućuje na neku stranu i znamo tekst linka, znamo i izlazne linkove te strane i tekstove linkova, nekad je to dovoljno za klasifikaciju

Konekcioni serveri

- Upiti vezani za graf veba
  - Ko upućuje na dokument čije je URL u upitu?
  - Na koga upućuje dokument čije je URL u upitu?
- Skladišti i omogućuje pretragu *inlinks* i *outlinks URL*-ova
- Korisni su za kontrolu i optimizaciju *crawling*-a, za analizu veb grafova, za link analizu koja se koristi za rangiranje rezultata
- Veb pretraživači imaju i ovo, i dosta pažnje tome posvećuju, da radi brzo, daje informacije koje im trebaju u *crawling*-u i indeksiranju i ne zauzima previše memorije
- Kompresiona tehnika za graf veba Boldi&Vigna, 2004 - link

Poreklo ideje

- Analiza citata naučnih radova - radovi koji citiraju iste radove su slične tematike
- Kasnije je analiza citata iskorišćena da se utvrdi rejting časopisa
- Jedan rad citira rad koji je objavljen u dobrom časopisu i koji ima dosta citata
- Jedan rad citira rad koji je objavljen u časopisu slabog rejtinga i rad nema citata
- Jedan rad citira rad koji ima iste autore - autocitat
- Da li prethodne tri vrste citata imaju istu težinu?

Linkovanje veb strana

- Veb i skup naučnih radova nije isto
- Na vebu mnogo više učesnika, različiti interesi i ciljevi
- Spam je daleko prisutniji - postoje i u nauci radovi koji su nastali kao šala, ali ih je beznačajno malo - Rad u časopisu Metalurgia International
- Od kada su se počeli linkovi uzimati u ubzir prilikom rangiranja rezultata u veb pretraživačima (1998) link spamovi rastu
  - Postoje i link farme - grupe veb sajtova koje se međusobno citiraju da bi dobili bolji rejting

Osnovna ideja

- Zamislimo da *web browser* se šeta *random* putanjom kroz veb sajtove
- Krene sa slučajno izabrane veb strane i dalje ide nekim od linkova koji se nalaze na toj strani, pri čemu je izbor linka slučajan i jednako verovatan - ako ima tri linka onda je težina za odlazak na svaku stranu 1/3
- U dugoj šetnji svaka veb strana ima težinu za posetu - *page score*

Teleportovanje

- Problem: veb graf nije jako povezan graf, nego je sačinjen od mnogo podgrafa koji su slabo povezani
- Rešenje: teleportovanje
  - Ako smo došli do mrtvog čvora teleportujemo se u proizvoljni čvor pri čemu je težina za prelaz jednaka 1/*ukupan broj čvorova*
  - U bilo kom čvoru (koji nije mrtav) se može teleportovati u proizvoljni čvor pri čemu je težina za prelaz jednaka α/*ukupan broj čvorova*, ili u čvor prema kojem postoji link pri čemu je težina za prelaz jednaka (1-α)/*broj izlaznih linkova*
  - α je parametar u intervalu [0,1] koji predstavlja verovatnoću da korisnik unese veb adresu direktno u polje za adresu (tipično ima vrednost 0.1)
- Više se kretanje ne može zaglaviti u podgrafu veba, odnosno zaista svaka veb strana ima težinu za posetu

Markovljevi lanci

- Markovljev lanac se sastoji iz n stanja i matrice prelaza nxn koja sadrži verovatnoće prelaza u drugo stanje
- Za svako $1 \leq i \leq n$ i $1 \leq j \leq n$, element matrice P_{ij} je verovatnoća da je j sledeće stanje ako se nalazimo u stanju i, pri čemu je $\forall i, \sum_{j=1}^{n} P_{ij} = 1$
- *Ergodic Markov chain* - iz proizvoljnog stanja **možemo** doći u bilo koje stanje, tako da u dugotrajnoj slobodnoj šetnji možemo obići sva stanja

Probabilistički vektori

- $x = (x_1, ..., x_n)$ - šetnja je u stanju i sa verovatnoćom x_i, pri čemu je $\forall i, \sum_{i=1}^{n} x_i = 1$
- Verovatnoće sa sledeće stanje su $x \times P$, $x \times P^2$, itd.
- Želimo verovatnoće da ste na nekom veb strani nezavisno kada i odakle ste krenuli u slobodnu šetnju po grafu
- $a = (a_1, ..., a_n)$ - vektor nepromenjivih verovatnoća
- $a = a \times P$, $a = a \times P^2$, itd.
- Računanje vektora a se svodi na računanje (levog) *eigenvector*-a matrice P
- a_i je *pagerank* stranice i

Zaključna razmatranja

- Pretprocerisanje - ne radimo pre svakog upita, nego periodično
  - Matricu P kreiramo upotrebom izlaznih linkova sa veb strana i unapred definisanog parametra α
  - Izračunamo vektor a koji predstavlja (levi) *eigenvector* matrice P
  - a_i je *pagerank* stranice i
- Procesiranje upita
  - Nađemo veb strane koje odgovaraju upitu
  - Rangiramo ih po *pagerank*
  - Problem: rangiranje nije zavisno od upita
  - Rešenje: to je samo jedan parametar u rangiranju

Osnove

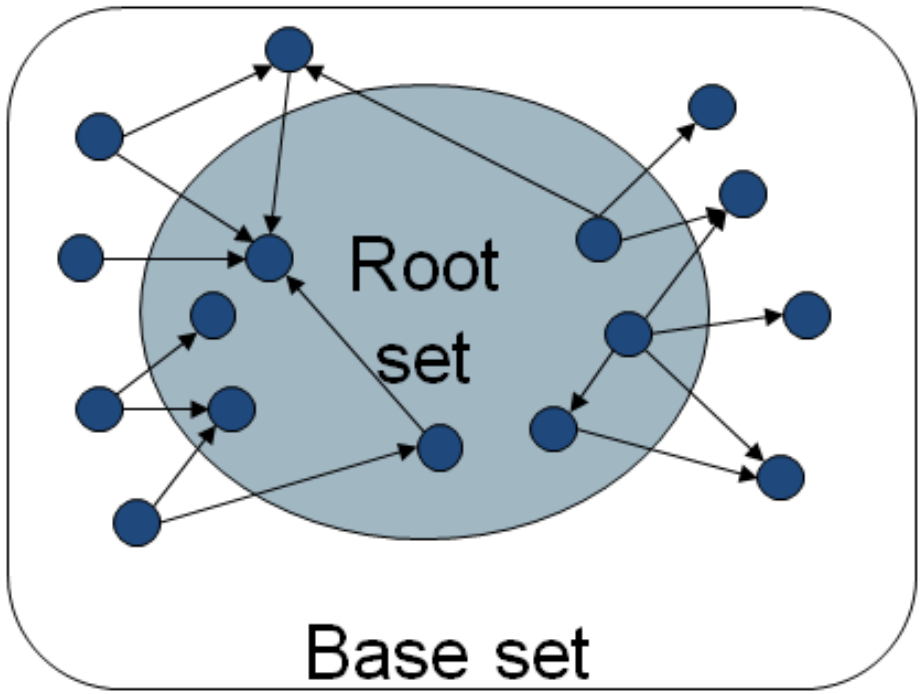
- *Hyperlink-Induced Topic Search*
- Dva skupa čvorova u veb grafu
  - *Hub* - čvorne tačke koje imaju puno linkova na veb strane iz odgovarajuće oblasti ("Bob's list of cancer-related links.")
  - *Authority* - značajan broj *Hub*-oba upućuje na ove strane koje se smatraju autoritetima u odgovarajućoj oblasti.
- Odlično za informacione potrebe koje nisu usko specificirane, nego se želi neko opšte mišljenje i informacije o nekoj oblasti - ovakve informacione potrebe nisu tako retko na vebu.

Cirkularna definicija

- Dobri čvorovi određene teme imaju puno linkova ka autoritativnim stranama
- Dobre autoritativne strane se spominju na puno čvorova
- Cirkularna definicija - računa se iterativno
- Odabere se početni skup čvorova i autoriteta
- Iterativno se skup smanjuje

Odabir početnog skupa

- *Root set* - sve što je odgovor na upit (npr. **browser**)
- *Base set* - *Root set* + *in-links* + *out-links*



Iteracija

- *hub score* - inicijalno za svako x je $h(x) = 1$
- *authority score* - inicijalno za svako x je $a(x) = 1$
- Nakon jedne iteracije je

$$h(x) = \sum_{x \to y} a(y) \tag{1}$$

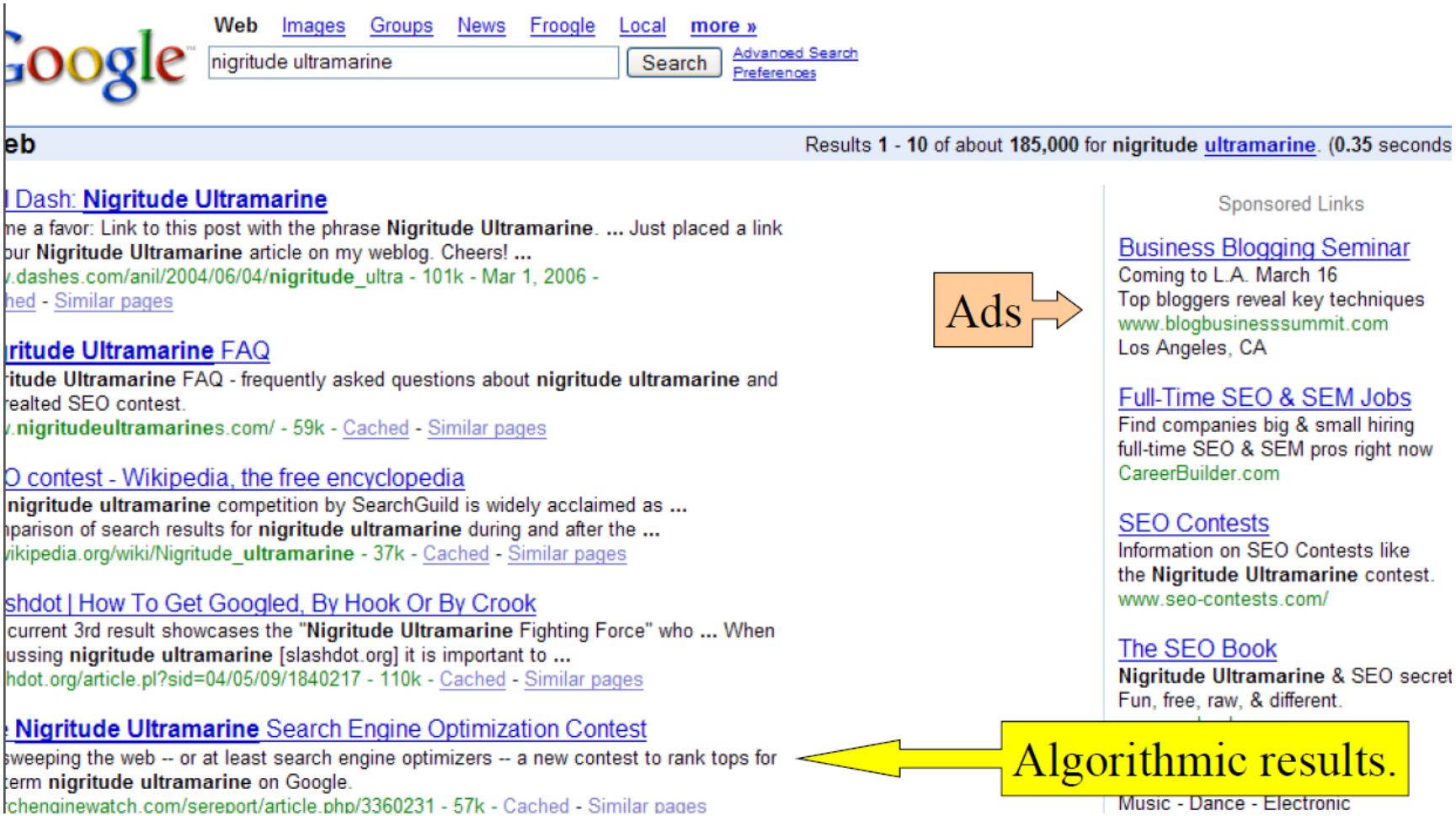
$$a(x) = \sum_{y \to x} h(y) \tag{2}$$

- Uzmemo samo top n čvorova i autoriteta i ponovimo iteraciju, ali ne vraćamo vrednosti na 1 (eventualno ih umanjimo k puta)
- U praksi 5 iteracija je dovoljno na čvorovi i autoriteti konvergiraju

Zapažanje

- Iteracije ne zavise od upita, samo osnovni skup
- Ako se u osnovnom skupu nađe stranica koja je imala malu relevantnost može uticati na autoritete i čvorove
- Povezani sajtovi sebi povećavaju *hub score* i *authority score*

Google ads


Google
Web Images Groups News Froogle Local more »
nigritude ultramarine
Search
Advanced Search
Preferences
eb
Results 1 - 10 of about 185,000 for nigritude ultramarine. (0.35 seconds
l Dash: Nigritude Ultramarine
me a favor: Link to this post with the phrase Nigritude Ultramarine. ... Just placed a link
our Nigritude Ultramarine article on my weblog. Cheers! ...
v.dashes.com/anil/2004/06/04/nigritude_ultra - 101k - Mar 1, 2006 -
hed - Similar pages
ritude Ultramarine FAQ
itude Ultramarine FAQ - frequently asked questions about nigritude ultramarine and
realted SEO contest.
v.nigritudeultramarines.com/ - 59k - Cached - Similar pages
O contest - Wikipedia, the free encyclopedia
nigritude ultramarine competition by SearchGuild is widely acclaimed as ...
nparison of search results for nigritude ultramarine during and after the ...
vikipedia.org/wiki/Nigritude_ultramarine - 37k - Cached - Similar pages
shdot | How To Get Googled, By Hook Or By Crook
current 3rd result showcases the "Nigritude Ultramarine Fighting Force" who ... When
ussing nigritude ultramarine [slashdot.org] it is important to ...
hdot.org/article.pl?sid=04/05/09/1840217 - 110k - Cached - Similar pages
Nigritude Ultramarine Search Engine Optimization Contest
sweeping the web -- or at least search engine optimizers -- a new contest to rank tops for
term nigritude ultramarine on Google.
rchenginewatch.com/sereport/article.php/3360231 - 57k - Cached - Similar pages
Ads
Sponsored Links
Business Blogging Seminar
Coming to L.A. March 16
Top bloggers reveal key techniques
www.blogbusinesssummit.com
Los Angeles, CA
Full-Time SEO & SEM Jobs
Find companies big & small hiring
full-time SEO & SEM pros right now
CareerBuilder.com
SEO Contests
Information on SEO Contests like
the Nigritude Ultramarine contest.
www.seo-contests.com/
The SEO Book
Nigritude Ultramarine & SEO secret
Fun, free, raw, & different.
Algorithmic results.
Music - Dance - Electronic

Google ads i *search results*

Google has maintained that ads (based on vendors bidding for keywords) do not affect vendors' rankings in search results

Search = ***miele***

Sponsored Links

CG Appliance Express
Discount Appliances (650) 756-3931
Same Day Certified Installation
www.cgappliance.com
San Francisco-Oakland-San Jose, CA

Miele Vacuum Cleaners
Miele Vacuums- Complete Selection
Free Shipping!
www.vacuums.com

Miele Vacuum Cleaners
Miele-Free Air shipping!
All models. Helpful advice.
www.best-vacuum.com

Web Results **1** - **10** of about **7,310,000** for **miele**. (**0.12** seconds)

Miele, Inc -- Anything else is a compromise
At the heart of your home, Appliances by **Miele**. ... USA. to **miele**.com. Residential Appliances. Vacuum Cleaners. Dishwashers. Cooking Appliances. Steam Oven. Coffee System ...
www.**miele**.com/ - 20k - Cached - Similar pages

Miele
Welcome to **Miele**, the home of the very best appliances and kitchens in the world.
www.**miele**.co.uk/ - 3k - Cached - Similar pages

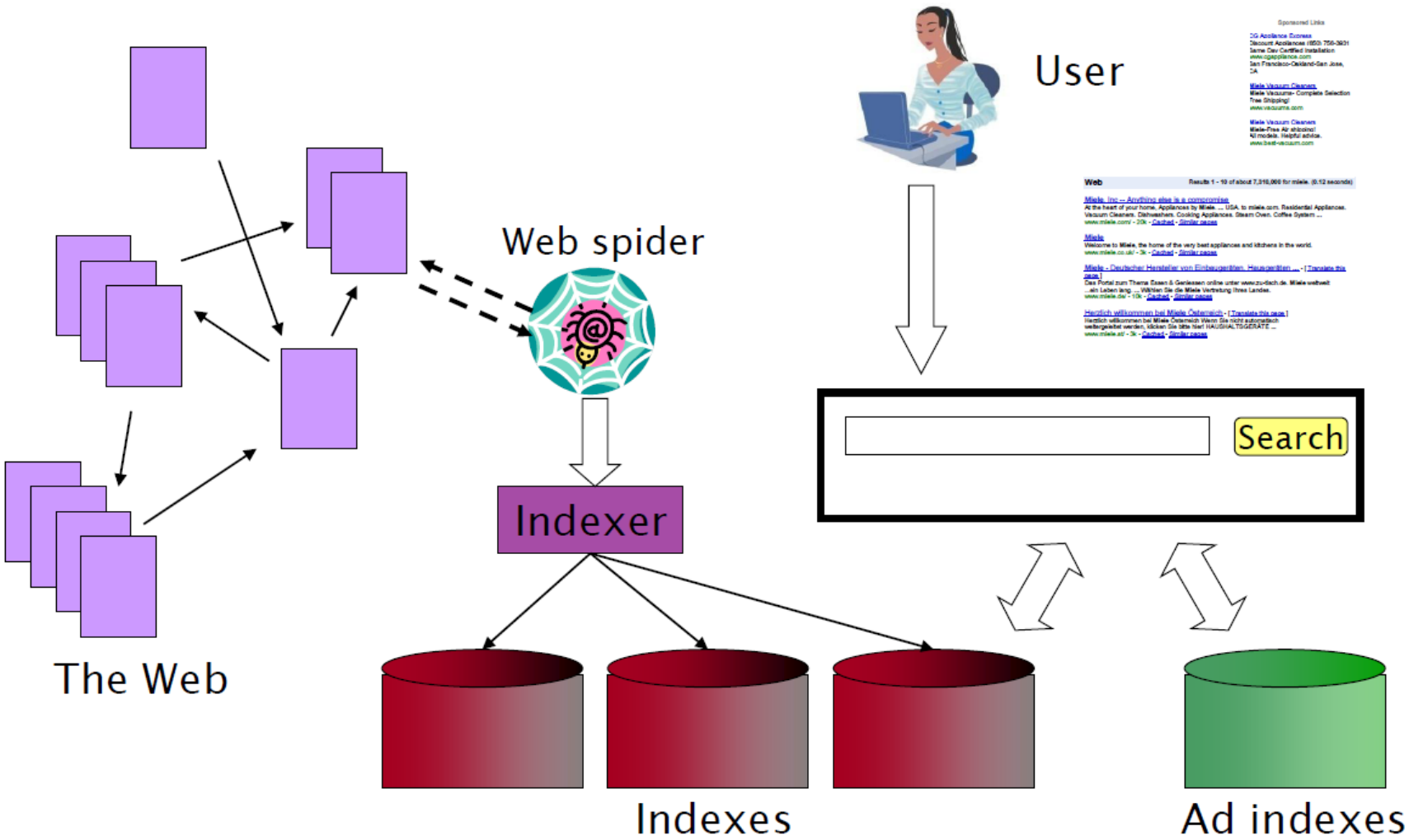
Miele - Deutscher Hersteller von Einbaugeräten, Hausgeräten ... - [Translate this page]
Das Portal zum Thema Essen & Geniessen online unter www.zu-tisch.de. **Miele** weltweit ...ein Leben lang. ... Wählen Sie die **Miele** Vertretung Ihres Landes.
www.**miele**.de/ - 10k - Cached - Similar pages

Herzlich willkommen bei **Miele** Österreich - [Translate this page]
Herzlich willkommen bei **Miele** Österreich Wenn Sie nicht automatisch weitergeleitet werden, klicken Sie bitte hier! HAUSHALTSGERÄTE ...
www.**miele**.at/ - 3k - Cached - Similar pages

Ads i *search results*

- Ostali veb pretraživači su usvojili ovaj koncept (Yahoo!, Bing)
  - Osnovna pretrage je nezavisna od plaćanja, ali postoji i prostor na ekranu za reklame, ovde je aukcija za ključne reči

Ads i *search results*



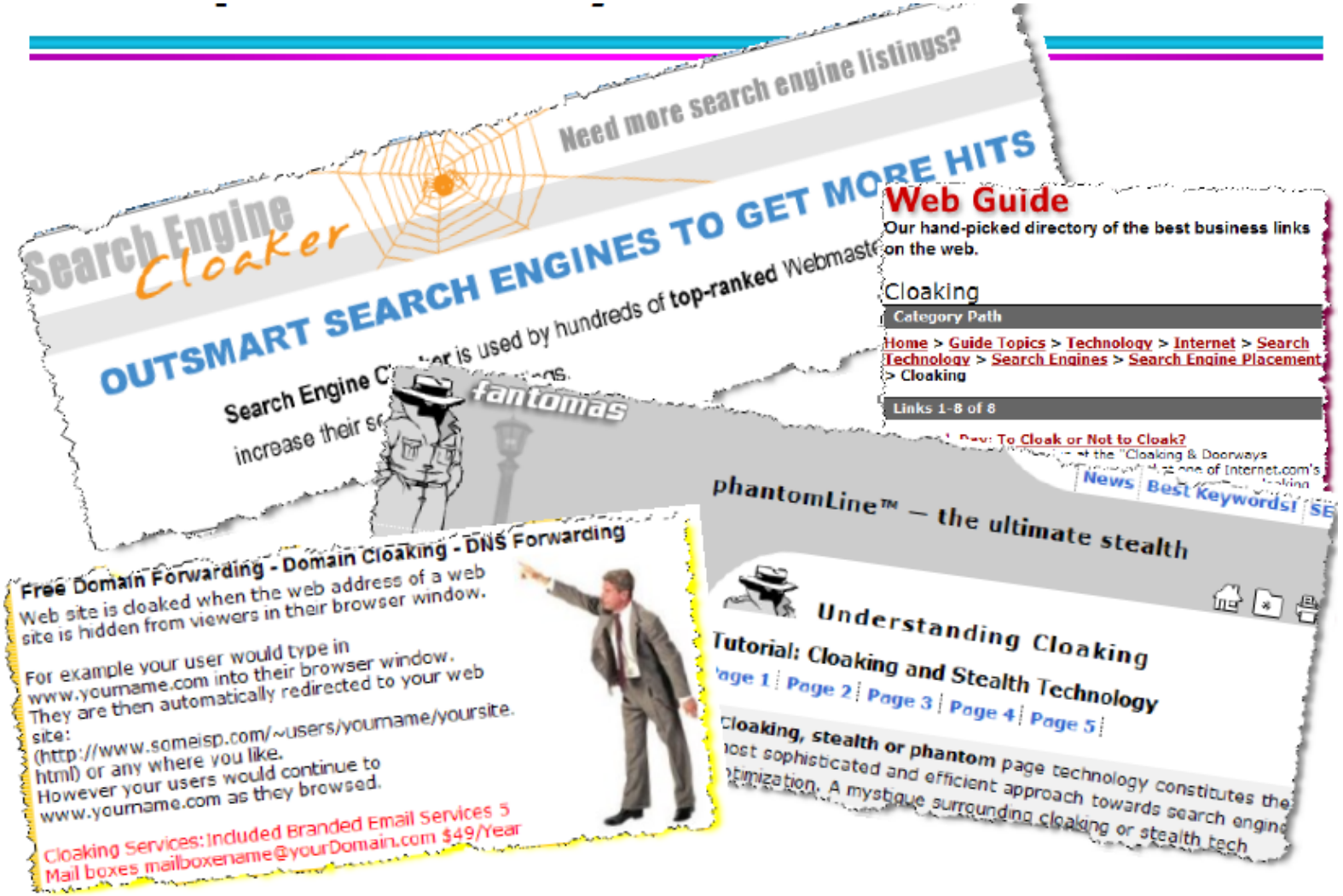
Search Engine Optimization

- Koji je problem sa plaćenim mestima za *Ads* - cena
- Koja je alternativa
- *Search Engine Optimization* (SEO)
  - Izmeni svoju veb stranu tako da ona bude dobro rangirana u algoritamskoj pretrazi (levi deo ekrana) za odgovarajuće ključne reči u upitu
  - Onda ne moraš da plaćaš veb pretraživačima ništa
  - Ali to ne znači da ne moraš ništa da platiš, to neko mora i da uradi
  - Suštinski ovo je deo marketinga kompanija
- SEO-om se bave kompanije, webmasters, konsultanti - zarada
- Neki se služe legitimnim sredstvima, a neki baš i ne - uputstvo, webmaster forum
- Ovo je spam ako dobijate kao najrelevantniji odgovor na Vaš upit nešto što nije najbolji odgovor

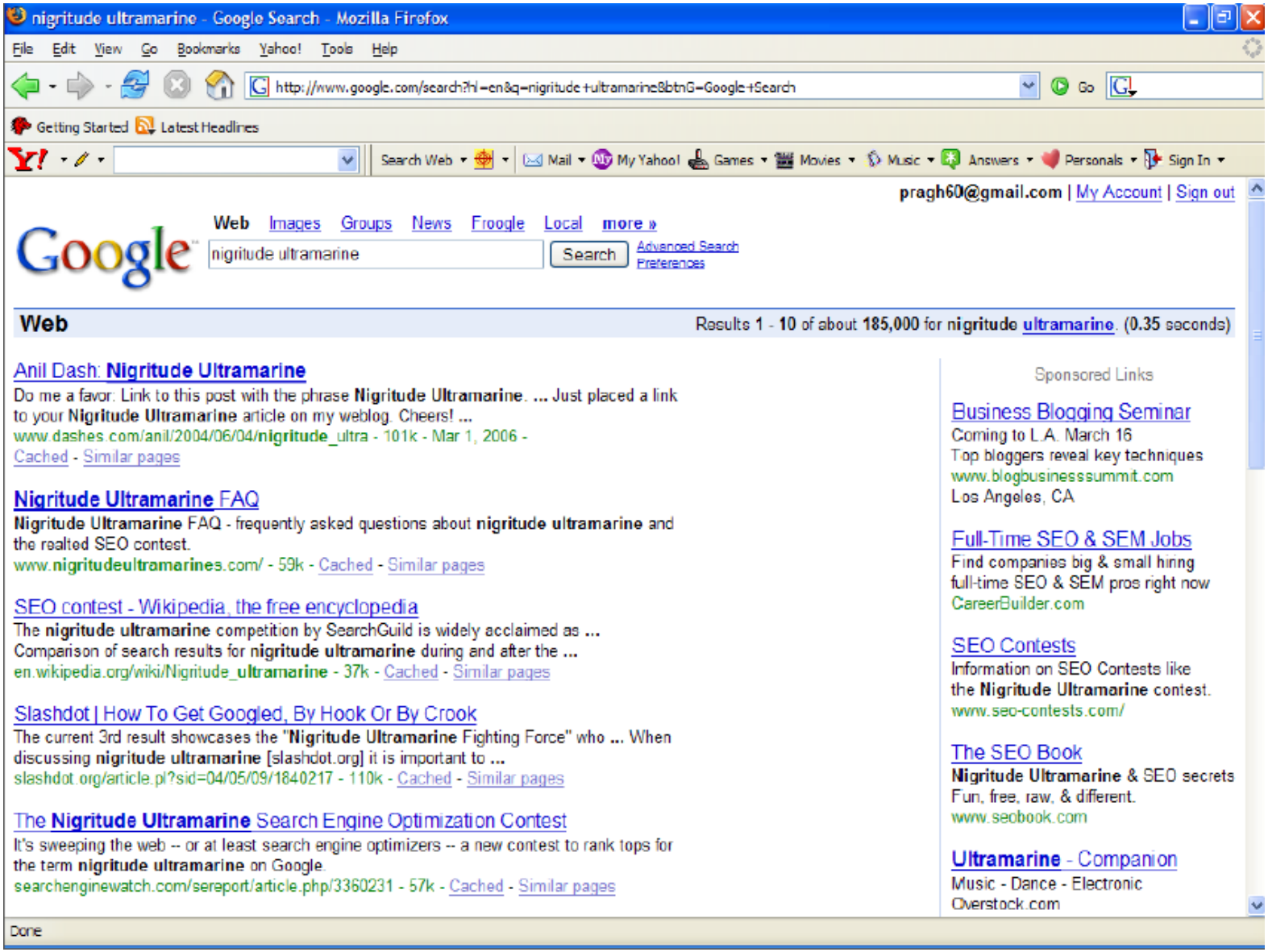
SEO i veb pretraživači

- Veb pretraživači obično imaju dokument u kojem navode šta je dozvoljeno
- Za dobronamerne to je polazna tačka šta da urade da povećaju vidljivost svoje stranice
- Za spamere ovo je polazna tačka da pronađu rupe opet sa istim ciljem da povećaju vidljivost svoje stranice a da ne plate, ili da šire svoje političke, religiozne ili druge ideje
- *Adversarial IR: the unending (technical) battle between SEO's and web search engines research* - link

SEO Industrija



SEO takmičenje

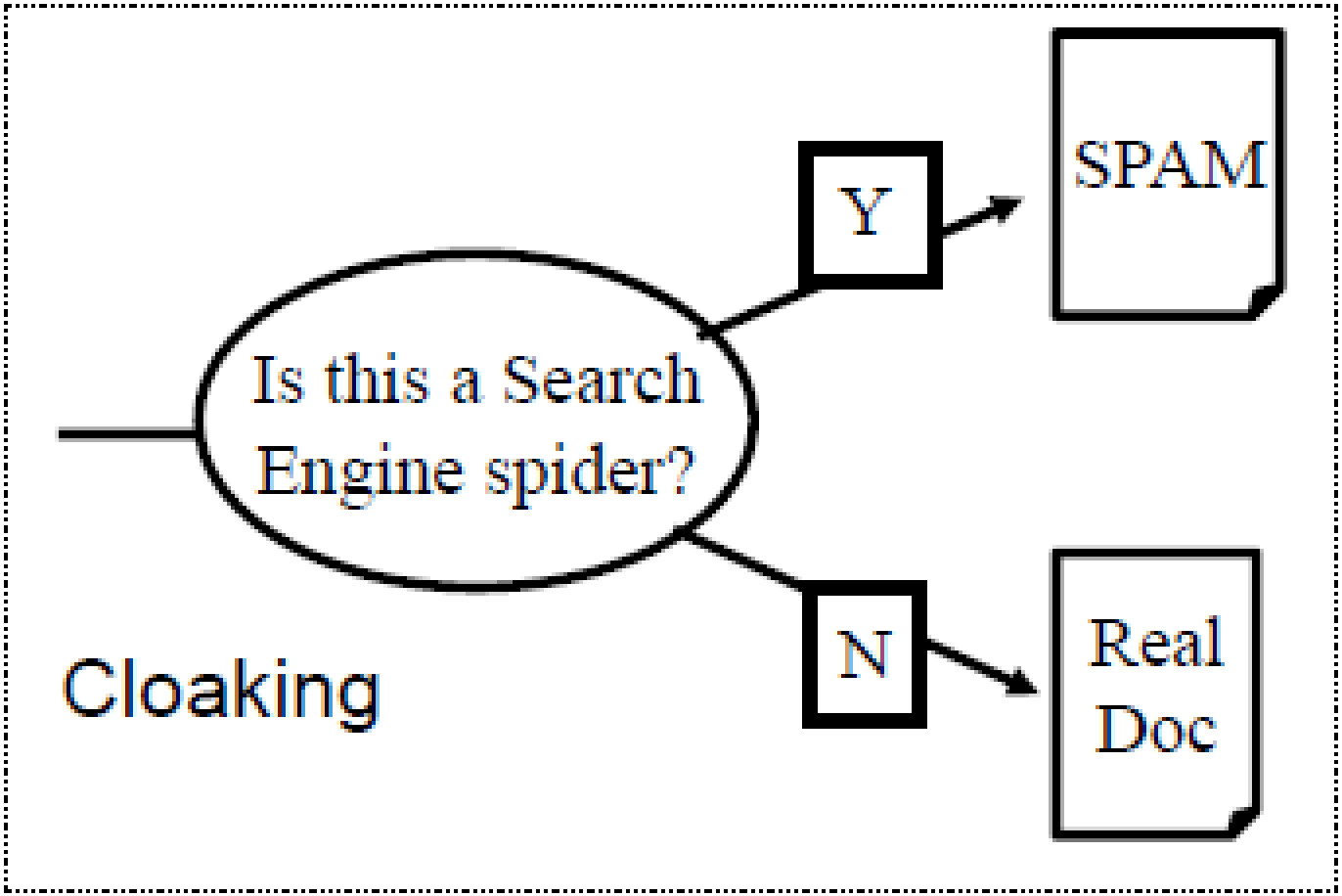


Jednostavne tehnike

- Prve generacije veb pretraživača su se oslanjale na tf/idf u rangiranju rezultata
- Top rangirane strane za upit "maui resort" su bile one koje su sadržale najviše reči "maui" i "resort"
- SEO tehnike su u html strane ponavljale kompletne sadržaje (ili bar bitne ključne reči) u boji pozadine
  - Ovi ponavljajući termovi su bili preuzeti i indeksirani od strane veb pretraživača, samim tim i uticali na tf/idf
  - Nisu bili vidljivi za ljude koji otvaraju stranice i nisu smetale čitaocima
- Trikovi sa *cascade style sheet*
- Namerno ponavljanje ključnih reči (ovo i nije toliko nelegitimno pravo)
- Netačne, ali popularne reči u meta-tags: *London hotels, hotel, holiday inn, hilton, discount, booking, reservation, sex, mp3, britney spears, viagra*

Cloaking

- Daje specijalno pripremljeni sadržaj *crawler*-ima



Ostale SEO tehnike

- *Doorway pages*
  - Stranice optimizovane za jednu ključnu reč od interesa i te stranice samo redirektuju do prave stranice
- *Link spamming*
  - Skriveni linkovi
  - Međusobno pomaganje - Link farme su grupe veb sajtova koje se međusobno citiraju
  - *Domain flooding* - gomila domena koji imaju link ili čak *redirect*-uju na ciljanu stranicu
- *Robots*
  - Milioni prijava putem *Add-Url*

Rat SEO vs veb pretraživači

- Pokušaj da se više gledaju linkovi sa autoritativnih strana, glasovi od autora, korisnika, itd.
- Link analiza da se proba detektovati spam
- Mašinsko učenje se koristi u detekciji spama, obučavajući skup su poznate spam stranice
- Detekcija veb strana koje se međusobno podržavaju (*family friendly filters*)
- *Blacklists*
- Žalbe
- *URL submission*-a uvodi anti robot testove
- Ograničenje ključnih reči u metapodacima