Univerzitet u Beogradu, Elektrotehnički fakultet

WEB PRETRAŽIVANJE

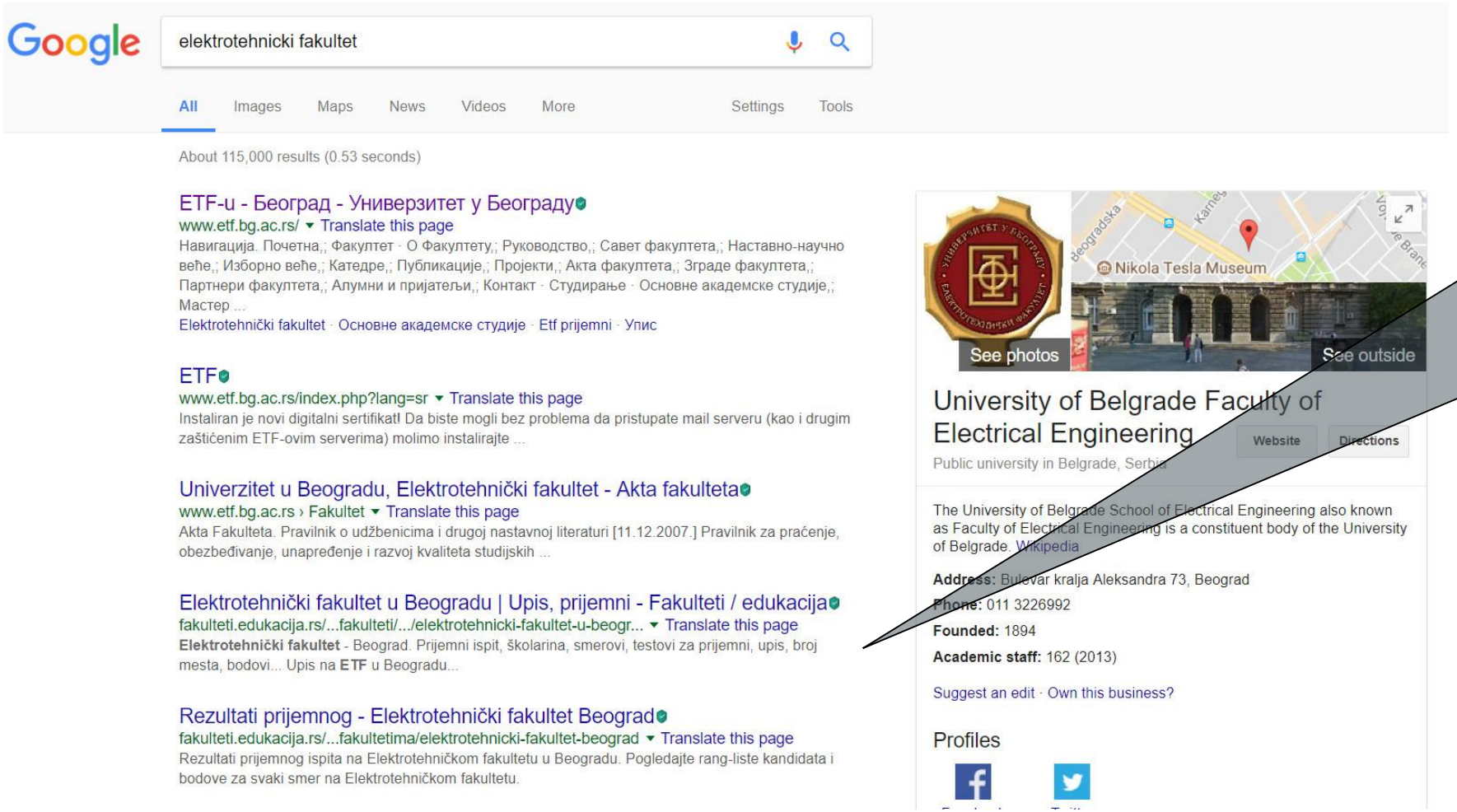
Miloš Pavković

Decembar, 2017

SADRŽAJ KURSA

- Motivacija, uvod u taksonomija i anatomiju Veb pretraživača.
- Problemi prilikom implementacije.
- Paralelizacija pretraživanja.
- Ponovno pretraživanje sajta.
- Specijalizovani/fokusirani pretraživači.
- Optimizacija pretraživanja od strane sajta.

UVOD



Q: Kako pretraživač zna da svi ovi sajtovi i strane sadrže traženu ključnu reč?
A: Zato što su sve te strane prvobitno indeksirane i parsirane uz pomoć Veb pretraživača

UVOD

- Definicija Veb pretraživača:
  - Veb pretraživač je kompjuterski program koji pretražuje internet mrežu na metodički, automatizovan način.
- Primena:
  - Prikupljanje stranica sa interneta.
  - Podrška globalnim pretraživačima kao što su Google, MSN i sl.
- Cilj:
  - Prikupiti sa stranica tekst, video, slike i sl.
  - Preko linkova i povezanosti strana rekonstruišu strukturu Veb sajta.

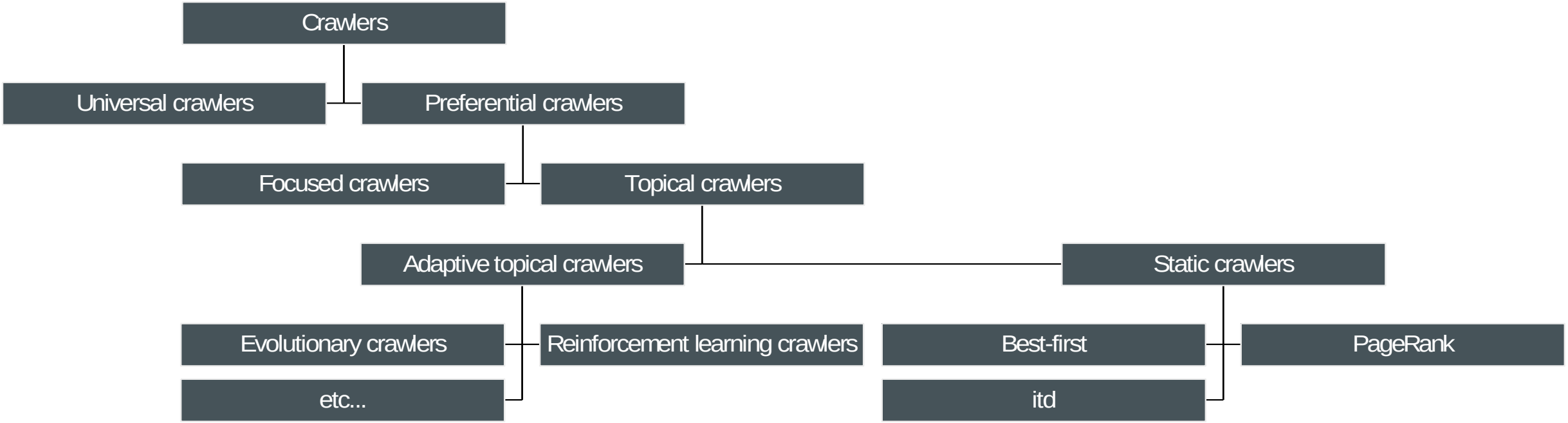
MOTIVACIJA

- U upotrebi kod svih značajnijih pretraživača (Google, Yahoo, MSN, Ask, itd.).
- Vertikalni (specijalizovani, fokusirani) pretraživači za vesti, kupovinu, dokumenta, korisnički generisan sadržaj itd.
- Biznis inteligencija, digitalno istraživanje javnog mnjenja, praćenje vremenskih nepogoda, epidemija (WHO).
- Praćenje Veb sajtova od interesa.
- Prikupljanje podataka (spamming, phishing…).

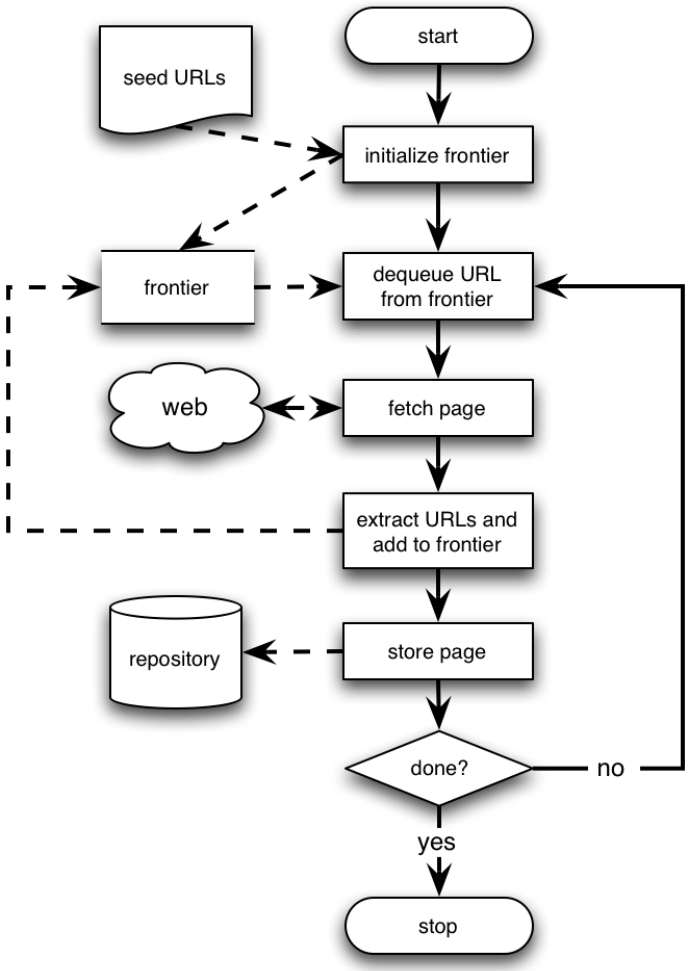
RAZLIČITI NAZIVI VEB PRETRAŽIVAČA

- Crawler
- Spider
- Robot (ili samo bot)
- Web agent
- Wanderer, worm, …
- I čuvene instance: googlebot, scooter, slurp, msnbot, …

TAXONOMIJA WEB PRETRAŽIVAČA



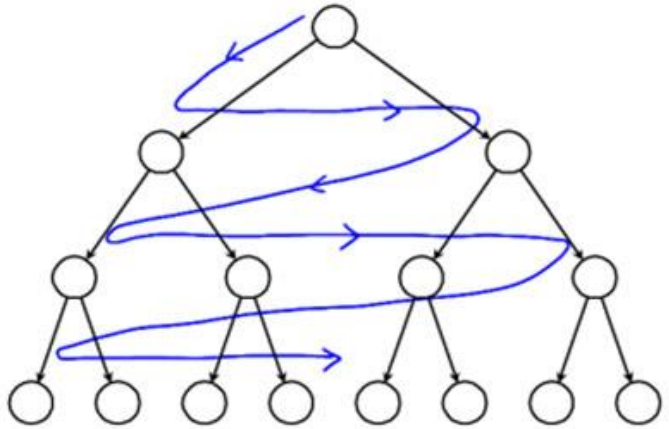
ANATOMIJA VEB PRETRAŽIVAČA



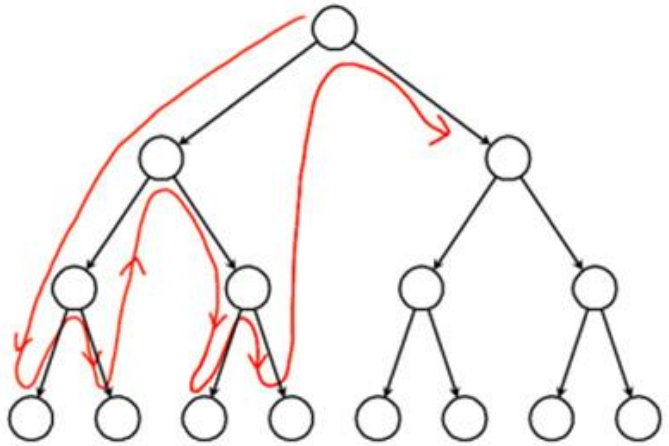
- Klasičan sekvencijalni pretraživač.
- SEED može biti bilo koja lista početnih URLa.
- Redosled posećivanja stranica određuje ‘frontier’.
- Svi novi nađeni URLovi se stavljaju u ‘frontier’ ako prethodno nisu bili posećeni.
- Posećene strane se smeštaju u repozitorijum.
- STOP kriterijum može biti bilo šta

PROLAZAK KROZ STRUKTURU WEB SAJTA (BFS ILI DFS)

- Obilazak po širini
  - Implementiran uz pomoć FIFO reda.
  - Pronalazi prvo strane najbliže ulazu sajta – najprioritetnije.
  - Ako za početak izaberu ‘dobre’ strane, velika je šanse da će ostale strane takođe biti ‘dobre’.



- Obilazak po dubini
  - Implementiran uz pomoć LIFO reda.
  - Uglavnom se koristi kod specijalizovanih pretraživača.
  - Može da odluta (lost in cyberspace).



PSEUDO KOD PRETRAŽIVAČA PO ŠIRINI NAPISAN U PERLU

```
my @frontier = read_seeds($file);
while (@frontier && $tot < $max) {
        my $next_link = shift @frontier;
        my $page = fetch($next_link);
        add_to_index($page);
        my @links = extract_links($page, $next_link);
        push @frontier, process(@links);
}
```

PROBLEMI PRILIKOM IMPLEMENTACIJE

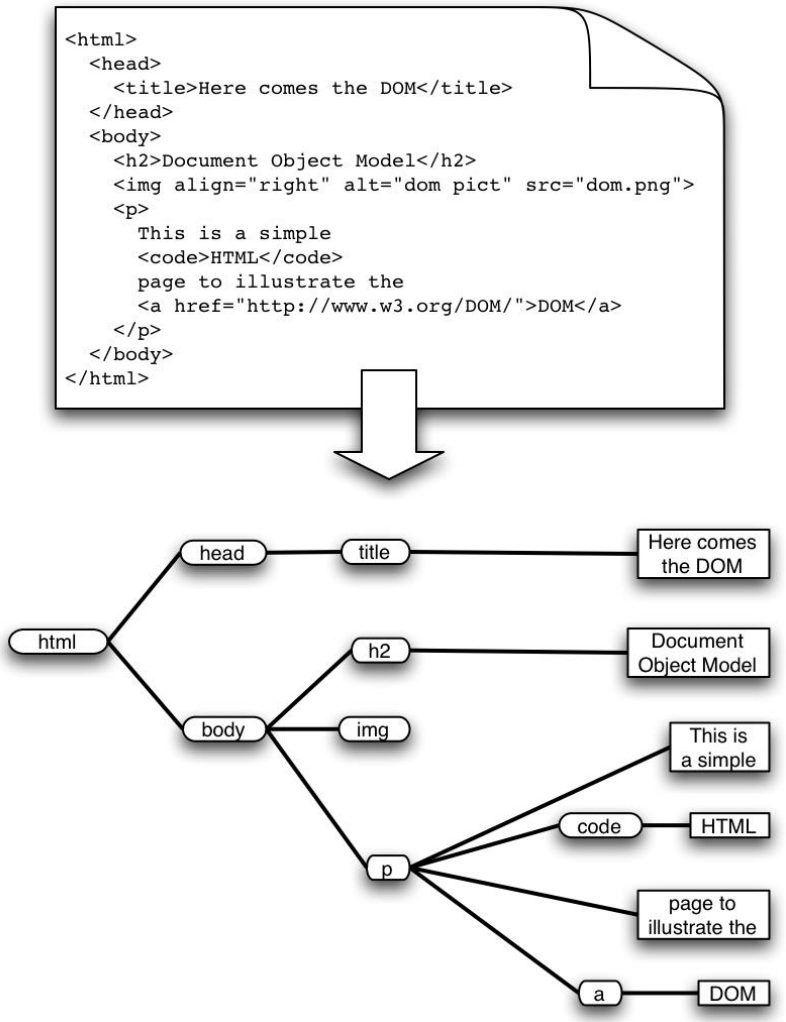
- Prikupljanje duplih stranica:
  - Čuvanje indeksa već prikupljenih stranica, što kroz tabelu što implementacijom HASH funkcija.
  - Takođe treba obratiti pažnju na strane koje još nisu posećene, a nalaze u frontieru.
- Rast frontiera:
  - Rešenje: prioritetizacija. Algoritmi koji daju prioritet jednim stranama u odnosu na druge.
- Modul za preuzimanje strana mora biti robustan:
  - Otporan na neuspelo preuzimanje strane, timeout mehanizam 404 greške i sl.
- Određivanje tipa fajla pre preuzimanja.

PROBLEMI PRILIKOM IMPLEMENTACIJE (PREUZIMANJE)

- Preuzimanje prvih 10KB do 250KB strane. Sve preko se naknadno preuzima ako ima potrebe.
- Menadžment *loop* linkova i linkova koji redirektuju. Po potrebi se ugrađuje *break* uslov.
- Obezbediti nastavak pretrage posle grešaka kao što su:
  - timeout
  - server not responding
  - file not found, i sl.

PROBLEMI PRILIKOM IMPLEMENTACIJE (PARSIRANJE)

- HTML je strukturno uređeno DOM stablo.
- Genirsano DOM HTML stablo na sajtu nije uvek tačno u smislu semantike.
- Pretraživači, isto kao i pregledači, moraju da ‘opraštaju’ greške.
- Postoje i gotovi alati koji vrlo dobro mogu da izađu na kraju sa lošim HTML kodom. Npr. - tidy.sourceforge.net.



PROBLEMI PRILIKOM IMPLEMENTACIJE (LINKOVI)

- Apsolutni VS. Relativni
  - Svaki pretraživač mora da prevede relativne linkove u apsolutne.
  - Pretraživač mora da dohvati 'Base URL' Veb sajta ili iz HTTP hedera, ili <base> taga zapisanog u HTML kodu.
  - Primer ako je base url: http://www.cnn.com/linkto/
    - Relative URL: intl.html
    - Absolute URL: http://www.cnn.com/linkto/intl.html
    - Relative URL: /US/
    - Absolute URL: http://www.cnn.com/US/

PROBLEMI PRILIKOM IMPLEMENTACIJE (URL KANONIZACIJA)

- Linkovi kao što su:
  - http://www.cnn.com/TECH
  - http://WWW.CNN.COM/TECH/
  - http://www.cnn.com:80/TECH/
  - http://www.cnn.com/bogus/../TECH/
- Svi odgovaraju jednoj kanonskoj formi: http://www.cnn.com/TECH/
- Da bi pretraživač izbegao duplikate svi linkovi moraju biti kanonizovani.
- Link koji sadrži port se pamti isključivo ako je port podrazumevani :80

PROBLEMI PRILIKOM IMPLEMENTACIJE (URL KANONIZACIJA)

- Neke transformacije su trivijalne ali se moraju poštovati:
  - ✗ http://informatics.indiana.edu
  - ✓ http://informatics.indiana.edu/
  - ✗ http://informatics.indiana.edu/index.html#fragment
  - ✓ http://informatics.indiana.edu/index.html
  - ✗ http://informatics.indiana.edu/dir1/./../dir2/
  - ✓ http://informatics.indiana.edu/dir2/
  - ✗ http://informatics.indiana.edu/%7Efil/
  - ✓ http://informatics.indiana.edu/~fil/
  - ✗ http://INFORMATICS.INDIANA.EDU/fil/
  - ✓ http://informatics.indiana.edu/fil/

OPTEREĆENJE

- Bitno je paziti kako se Veb sajt pretražuje u potrazi za stranama.
- Cloudflare zaštita (CDN, DNS, DDoS) – sajt će vas blokirati ako ga preopteretite.
- Nasumično vreme preuzimanja strane.
- Slučajno izabrane strane.
- Nasumične strane na nasumičnih sajtova.
- Promena IP adresa odakle se vrši pretraga - VPN, TOR.
- Automatsko logovanje i simulacija korisnika (nije preporučljivo).

PARALELIZACIJA - MOTIVACIJA

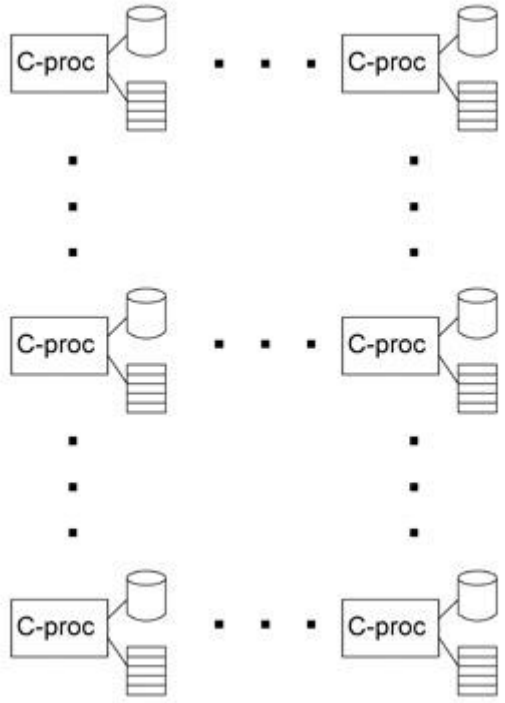
- Skalabilnost – zbog velikog sadržaja interneta pa i samih sajtova, od imperativa je paralelizovati pretraživanje i ubrzati proces prikupljanja. Više procesa može da pretražuje jedan sajt.
- Disperzija opterećenja mreže – procesi mogu da se izvršavaju na geografski različitim lokacijama.
- Redukcija opterećenja mreže – pretraživač na jednoj geografskoj lokaciji ne treba da pretražuje udaljene sajtove. Na ovaj način se ne opterećuje mreža kroz koje prolaze zahtevi.

PARALELIZACIJA - PROBLEMI

- Preklapanje – različiti procesi mogu da prikupe iste strane. Jedan proces mora biti svestan šta su drugi prikupili u međuvremenu.
- Kvalitet – odluka o prioritetnim stranama se teže donosi. Jedan proces ne može biti na pravi način svestan ‘slike’ celog sajta.
- Opterećenje intra-proces mreže – da bi se poboljšao kvalitet strana i sprečilo preklapanje, procesi moraju da komuniciraju između sebe. Što veći broj procesa i strana koje se preuzimaju, veće je opterećenje mreže kroz koji procesi komuniciraju.

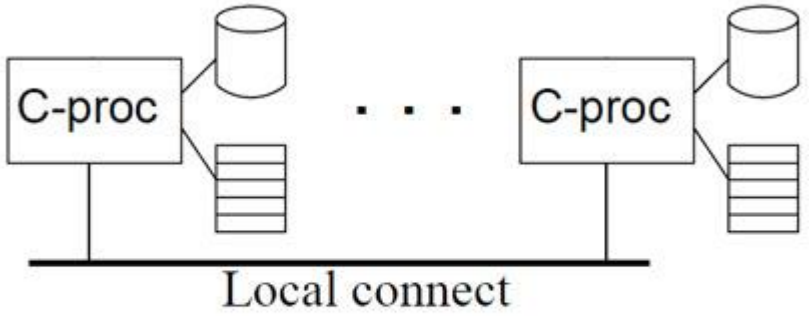
PARALELIZACIJA – ARHITEKTURA

- Paralelni pretraživač mora da se sastoji od više procesa koji obavljaju identični posao
- Svaki proces treba da preuzme stranu sa Veb sajta, ekstraktuje linkove, i skladišti preuzetu stranu u bazu.
- U zavisnosti kako se preuzeti linkovi raspoređuju kasnije između procesa, razlikujemo intra-site paralelne pretraživače i distribuirane pretraživače.



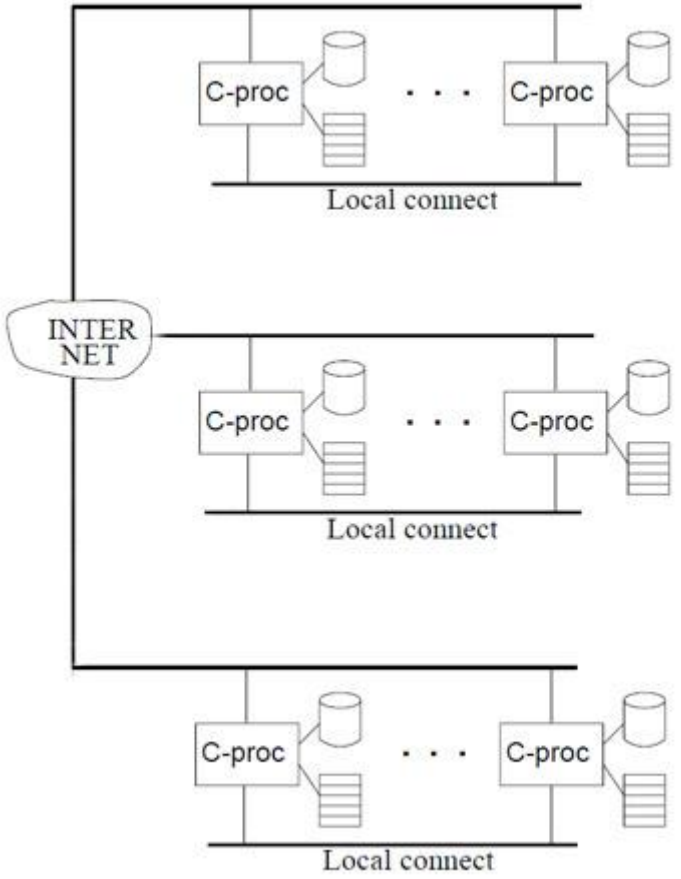
PARALELIZACIJA – INTRA-SITE PRETRAŽIVAČ

- Kada svi procesi rade na jednoj lokalnoj mreži.
- Komunikacija se obavlja preko brze lokalne infrastrukture.
- U ovom slučaju opterećenje mreže prilikom razmene informacija je centralizovano.



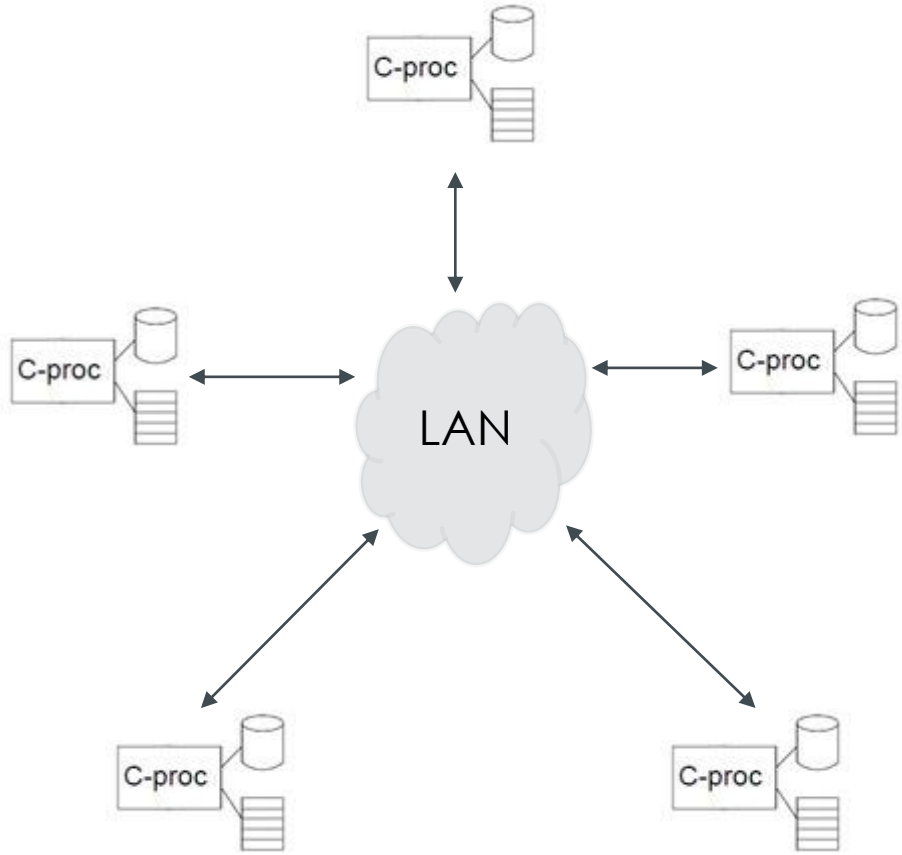
PARALELIZACIJA – DISTRIBUIRANI PRETRAŽIVAČ

- Kada svi procesi rade na geografski udaljenim lokacijama i povezani su preko interneta.
- Ovakvi distribuirani potraživači pomažu pri rasterećenju mreže i disperziji pretraživanja.
- Problem prilikom odluke kada treba komunicirati i kojom frekvencijom.



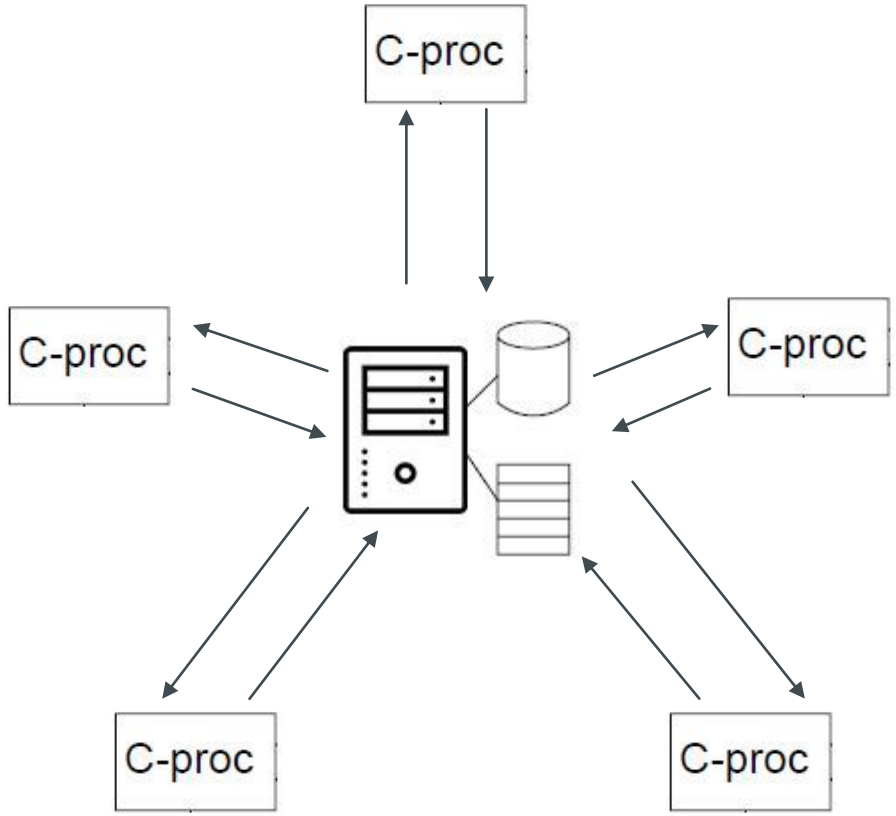
PARALELIZACIJA – NEZAVISNI PROCESI

- Svaki proces sam odlučuju koje strane će da preuzme.
- Ovo može biti urađeno sa ili bez konsultacije sa ostalim procesima.
- Što manje koordinacije, veći broj preklapanja ali znatno rasterećenija mreža.
- Veća komunikacija opterećuje procese pa samim tim i usporava pretragu.
- Potrebno je balansirati između brzine i kvaliteta prikupljanih podataka.



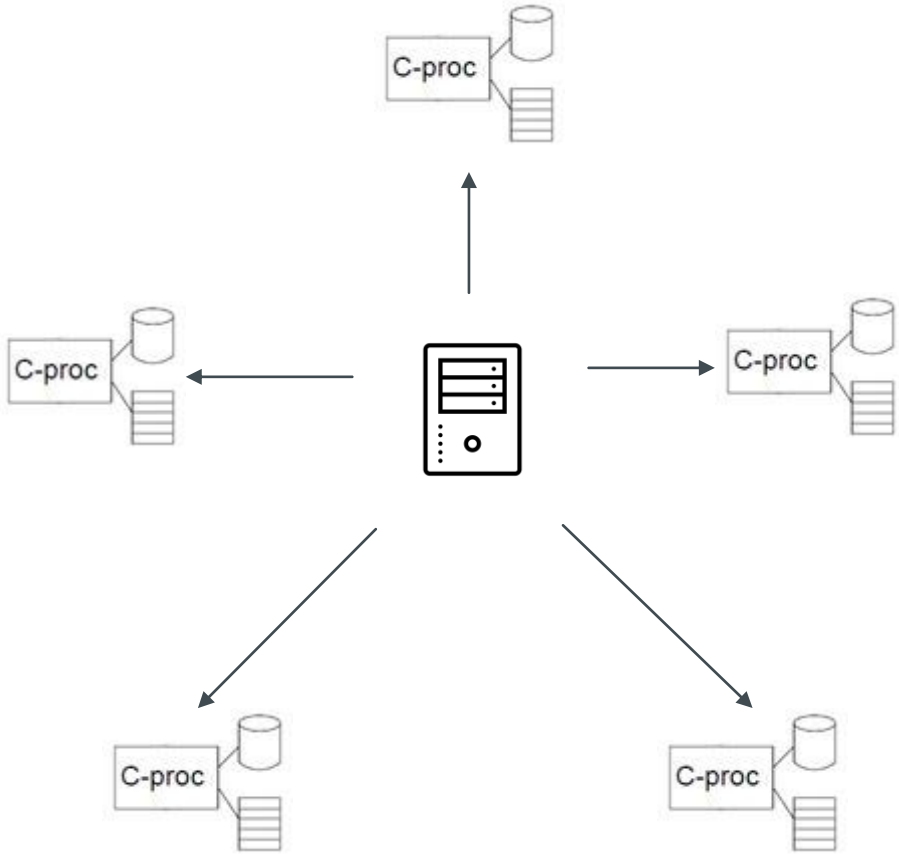
PARALELIZACIJA – CENTRALIZOVANA KOORDINACIJA

- Takođe se naziva i dinamički mod.
- Strane su logički particionisane i raspoređene procesima od strane glavnog procesa-servera.
- Procesi parsiraju linkove sa preuzete strane i vraćaju podatke glavnom koordinatoru.
- Sporije pretraživanje zbog frekventne komunikacije sa glavnim koordinatorom.
- Smanjen broj duplih stranica.



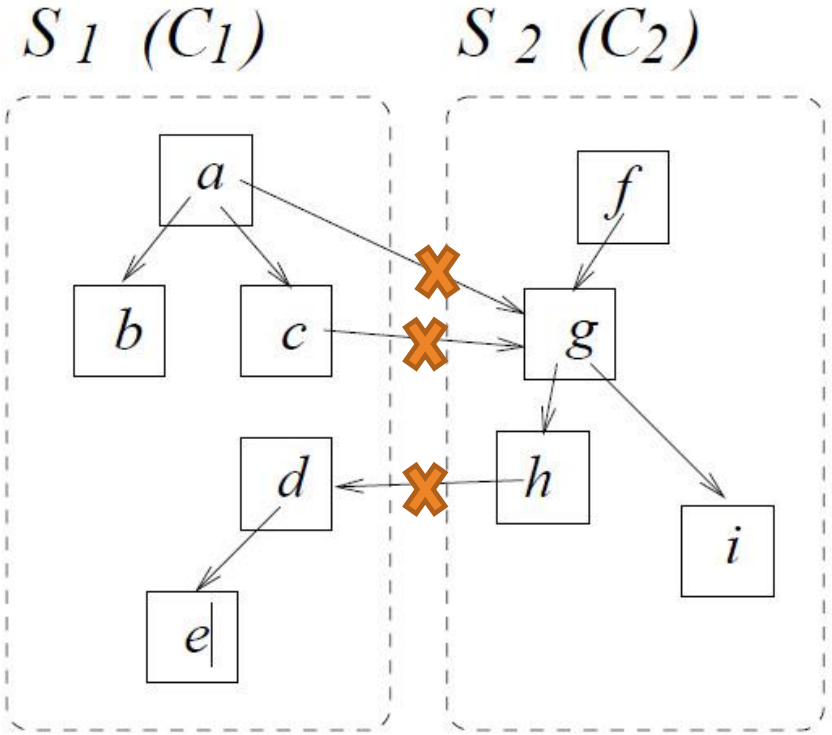
PARALELIZACIJA – STATIČKI MOD RADA

- Procesima se samo u startu podele logički particionisani delovi sajta ili veb prostora koji treba da se pretraži.
- Proc1: http://informatics.indiana.edu/dir1/ Proc2: http://informatics.indiana.edu/dir2/
- Nema potrebe za centralnim koordinatorom i međusobnom koordinacijom sem na početku.
- Problem sa linkovima koji vode u domen drugih procesa.



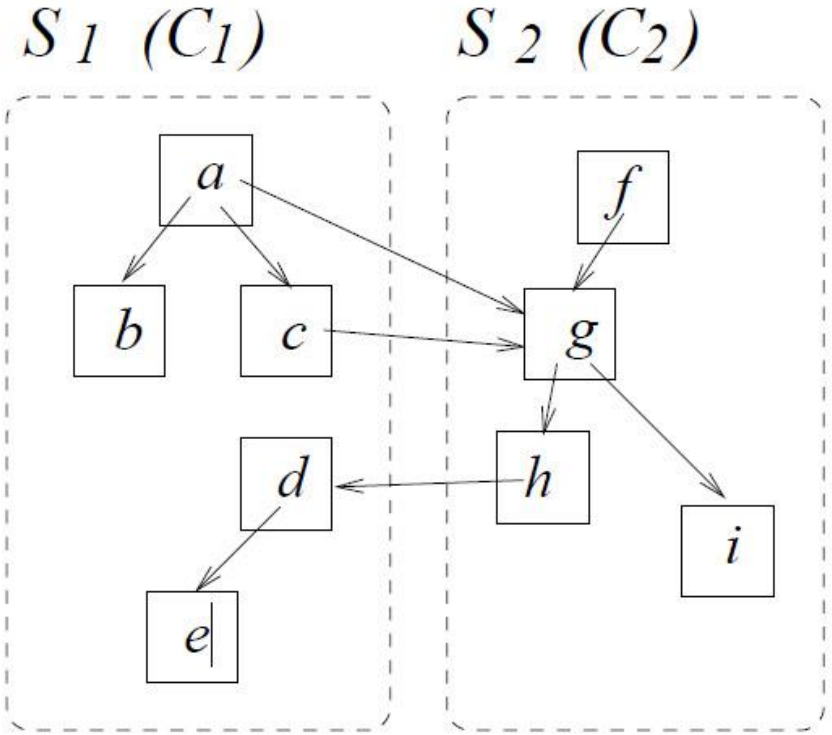
PARALELIZACIJA – STATIČKI MOD RADA (FIREWALL MODE)

- Procesi ignorišu linkove koji vode van njihovih particionisanih domena.
- linkovi $c \rightarrow g$, $a \rightarrow g$, $h \rightarrow d$ se ignorišu pa samim tim nema potrebe za koordinatorom.
- Nema preklapanja ni opterećenja.
- Strane kao što su *e* i *d* nije moguće dohvatiti iz istog domena, pa ostaju ne pretražene.



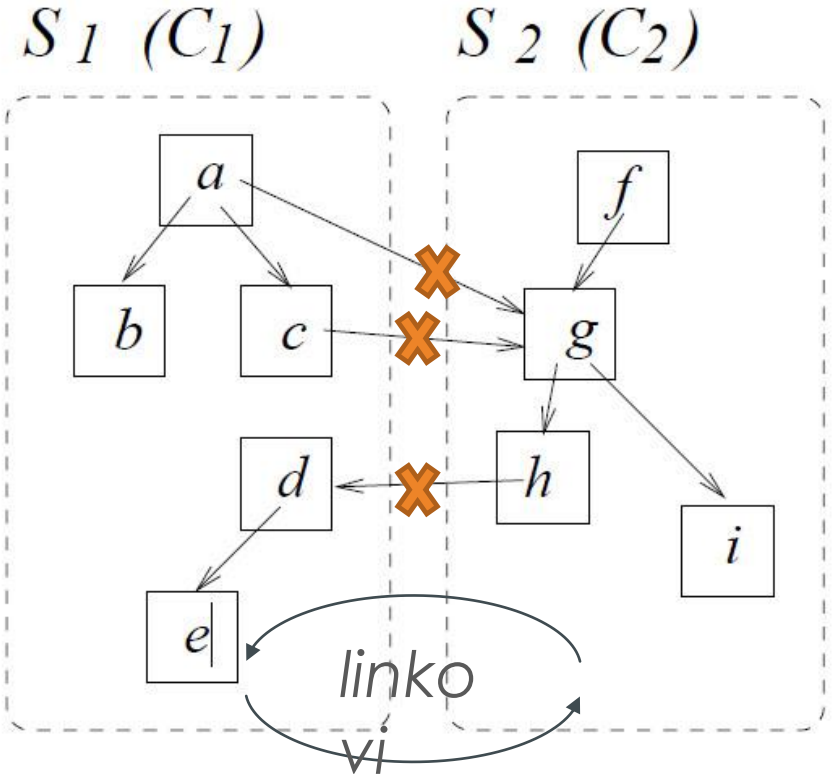
PARALELIZACIJA – STATIČKI MOD RADA (CROSS-OVER MODE)

- Procesi prvo obiđu sve strane svog domena, pa tek onda izlaze van njega.
- Na ovaj način proces *C1* može da dohvati strane *d* i *e* preko domena procesa *C2*.
- Postoji preklapanje i neke strane se preuzimaju duplo.
- Ograničavanje procesa na koliko drugih domena je dozvoljeni preći i koliko strana van svog domena sme da pretraži pre nego što se zaustavi.



PARALELIZACIJA – STATIČKI MOD RADA (EXCHANGE MODE)

- Procesi periodično razmenjuju informacije sa drugim procesima o linkovima koji vode na njihov domen, ali ih ne prate.
- Proces *C1* informiše proces C2 o postojanju linka $c \rightarrow g$ i $a \rightarrow g$. *C2* informiše *C1* o $h \rightarrow d$.
- Sve strane domena procesa bivaju obiđene.
- Minimalna komunikacija između procesa i minimalno opterećenje mreže.



PONOVNO PRETRAŽIVANJE SAJTA

- Posle nekog vremena potrebno je posetiti sajt u potrazi za novim sadržajem.
- Sadržaj na starim stranama se možda izmenio.
- Neke strane više ne postoje.
- Nova strane su dodate i potrebno ih je pronaći.

PONOVNO PRETRAŽIVANJE SAJTA

- Postoje dva moda pretraživanja sajta. Inicijalno i ponovno.
- Inicijalno pretraživanje - kada se sajt posećuje prvi put i sve strane se indeksiraju.
- ponovno pretraživanje - kada se posle određenog vremena ponovo posećuju strane sajta u potrazi za novim ili izmenjenim sadržajem.
- Problem – frekvencija posete i procena izmene na stranama.

NAČINI PONOVNOG PRETRAŽIVANJA SAJTA

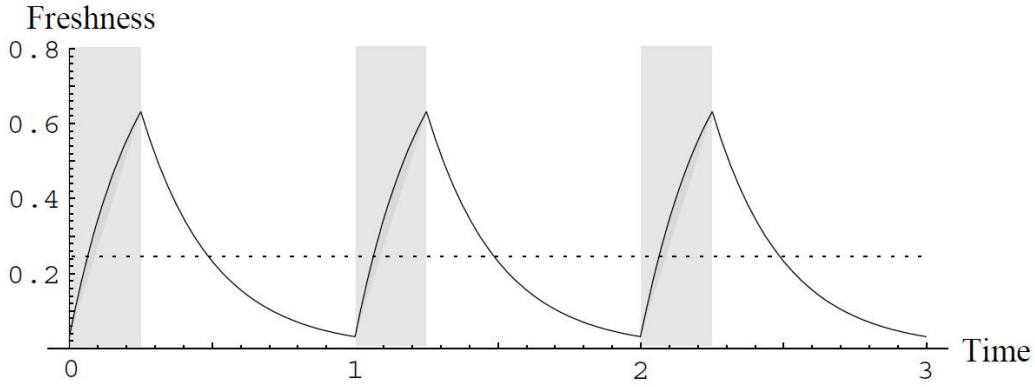
- Periodični i Inkrementalni.
- Periodični pretraživač posle određenog vremena ponovo obilazi kompletan sajt i stare strane zamenjuje novim.
- Inkrementalni konstanto posećuje bitne strane i ažurira bazu - '*Batch-mode*' i '*Steady*' modovi pretraživanja.
- Inkrementalni su bolji u situaciji gde se samo određene strane menjaju. Ovo doprinosi boljem čuvanju protoka i bržem otkrivanju novih strana.
- Periodični pretraživač otkriva nove strane tek kada poseti ponovo ceo sajt, što u većini slučajeva nije efikasno.

BALANSIRANJE IZMEĐU POKRIVENOSTI I SVEŽINE

- Coverage & Freshness – dva najbitnija faktora pri inkrementalnom pretraživanju.
- Proceniti životni vek strane i sadržaja na strani. Koliko se često sadržaj menja. Koliko često se dodaju nove stranice a stare brišu.
- Veća pokrivenost i traženje novih strana smanjuje broj poseta postojećim stranama.
- Većom pokrivenošću se smanjuje svežina sadržaja.
- Inkrementalni pretraživač mora da balansira između pokrivenosti i svežine.

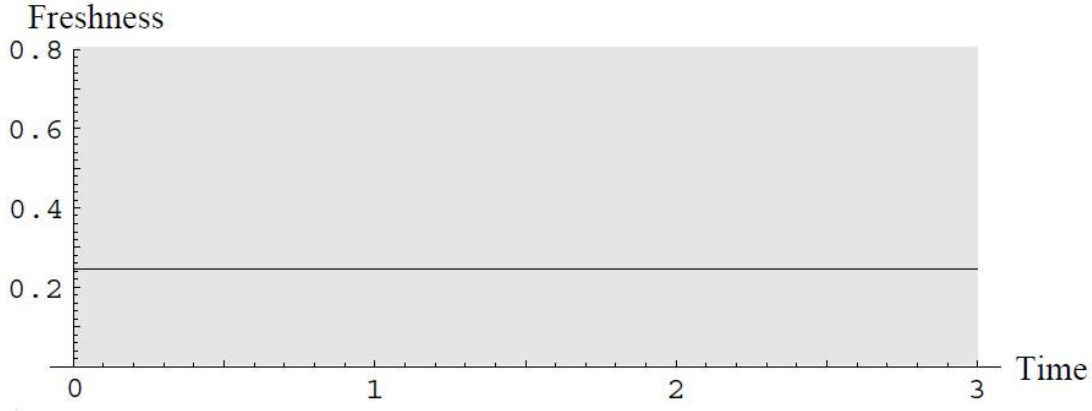
‘BATCH-MODE’ INKREMENTALNI PRETRAŽIVAČ

- Batch-mode pretraživač periodično posećuje strane u određenim vremenskim intervalima.
- Kada procenjena vrednost svežine baze opadne, inicijalizuje se nova pretraga.



‘‘STEADY’ INKREMENTALNI PRETRAŽIVAČ.

- ‘Steady’ pretraživač konstantno posećuje sajt.
- Nove stane se otkrivaju i stari sadržaj se zamenjuje novim u bazi.
- Prosečna svežina stranica je ista kao i kod ‘batch’ modela.
- ‘Steady’ model ima prednost jer posećuje sajt manjom brzinom i samim tim manje opterećuje mrežu.



PRIKUPLJANJE NOVOG SADRŽAJA

- Shadowing i In-Place.
- Shadowing – tehnika gde se sav novi sadržaj prvo sakupi odvojeno, pa se na kraju kopira preko starog od jednom.
- In-Place – tehnika gde preko starog sadržaja odmah presnimi novi čim se na njega naiđe.
- Shadowing tehnika omogućava da se trenutni sadržaj zaštiti o samog procesa pretraživanja i trenutnih nepravilnosti koji mogu da se dese kod povezanih informacija između strana. Odgovara najviše ‘batch’ modu.
- In-Place tehnika omogućuje da se novi sadržaj što pre nađe u bazi i samim tim povećava svežinu. Odgovara najviše ‘steady’ modu.

FREKVENCIJA POSETE STRANAMA

- Fiksirana i varijabilna frekvencija.
- Kod fiksirane frekvencije pretraživač posećuje sve strane u isto vreme u potrazi za novim sadržajem. Najviše odgovara ‘batch’ modelu.
- Kod varijabilne frekvencije pretraživač za svaku stranu nezavisno određuje vreme ponovne posete. Najviše podobna za ‘steady’ model.
- Različiti matematički modeli za procenu frekvencije promene sadržaja na strani.

DVE MOGUĆE VARIJANTE INKREMENTALNOG PRETRAŽIVAČA

Tip A

Steady
In-Place
update
Variable freq.

- Sveži podaci
- Manja opterećenost mreže

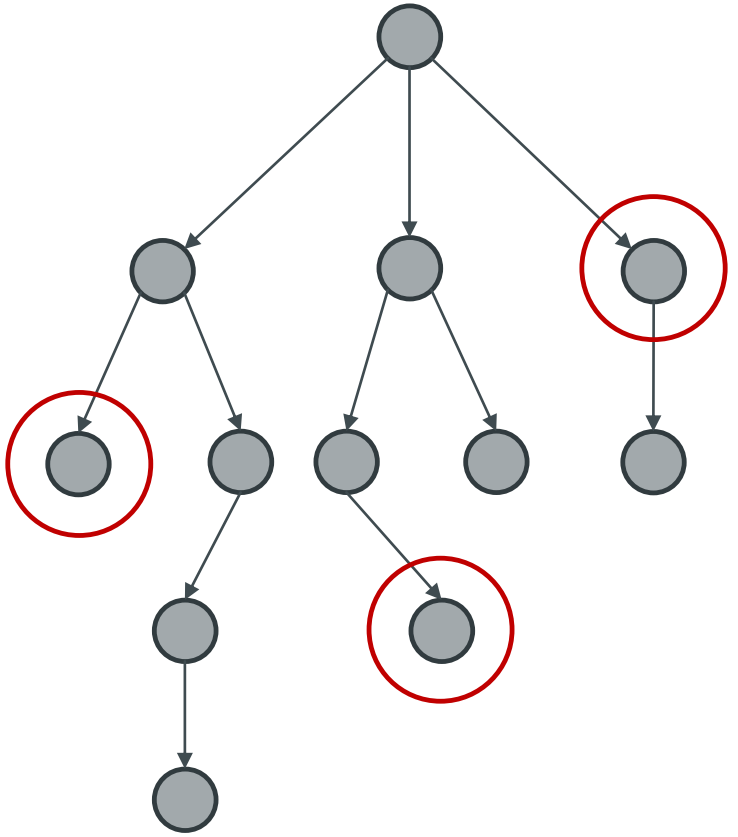
Tip B

Batch-Mode
Shadowing
Fixed freq.

- Lakši za implementaciju
- Visok integritet podataka u celini

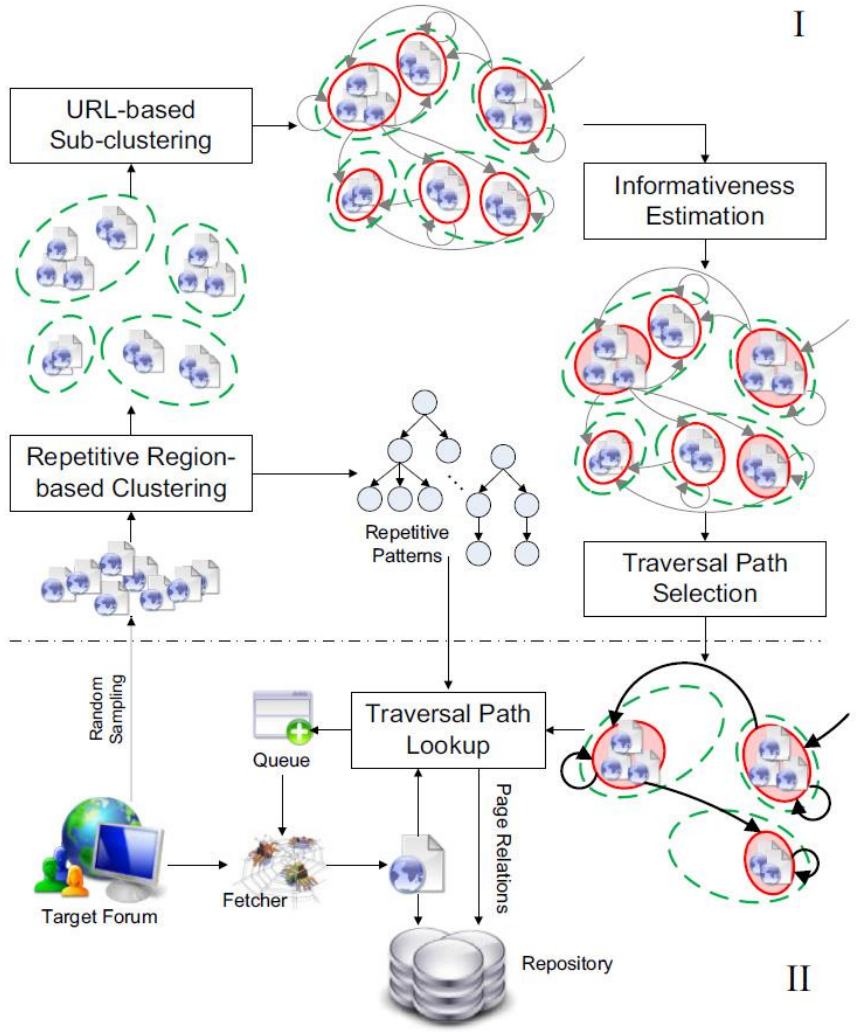
SPECIJALIZOVANI PRETRAŽIVAČI

- Takođe se nazivaju i fokusirani.
- Specijalno dizajnirani da traže samo određene strane sajta.
- Takođe mogu biti fokusirani na određeni sadržaj.
- Specijalizovan potraživač donosi odluke o daljem pretraživanju prepoznavanjem specifičnih linkova ili tipova stranica.
- Najpoznatiji su iRobot i FoCUS.



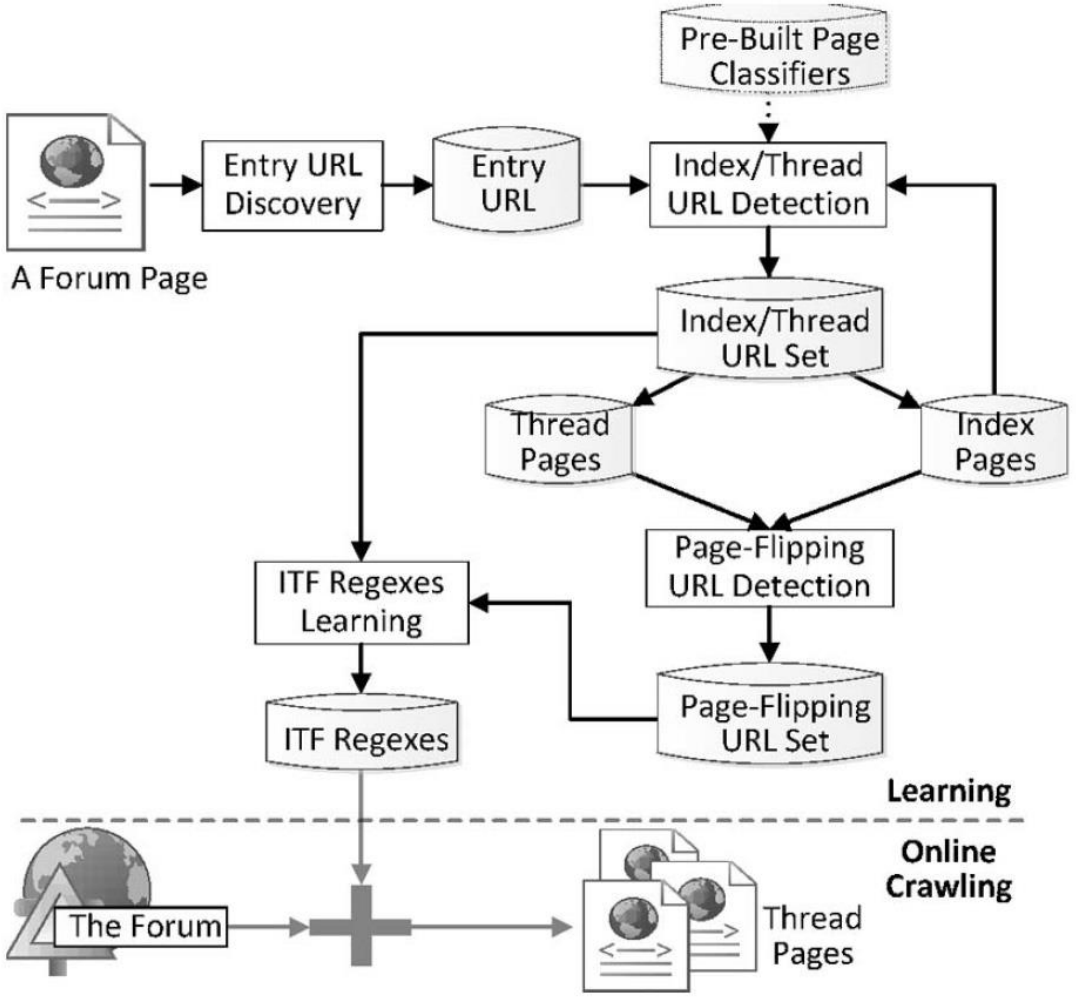
SPECIJALIZOVANI POTRAŽIVAČ BAZIRAN NA PREPOZNAVANJU STRANICA

- Najpoznatiji - iRobot.
- Sastoji se od dve faze, učenja i pretraživanja.
- U prvoj fazi se strane klasterizuju i izvlače parterni na osnovu kojih će se kasnije prepoznavati strane za vreme pretraživanja.
- U drugoj fazi se pretražuje sajt na osnovu šablona određenih u prvoj fazi.
- U zavisnosti šta je cilj pretraživanja, strane sa takvim sadržajem se ciljaju u prvoj fazi učenja.



SPECIJALIZOVANI POTRAŽIVAČ BAZIRAN NA PREPOZNAVANJU LINKOVA

- Najpoznatiji – FoCUS.
- U prvoj fazi se uz pomoć mašinskog učenja i SVMa uče šabloni o stranama.
- Uz pomoć naučenih šablona prepoznaju se strane i linkovi koji vode do njih.
- Učenje regularnih izraza na osnovu izabranih linkova.
- Druga faza je faza pretrage uz pomoć regularnih izraza koji prepoznaju linkove.



OPTIMIZACIJA PRETRAŽIVANJA SA STRANE SAJTA

- Robots.txt – tekstualna datoteka obično održavana od strane administratora ili tehnologije koja pokreće Veb sajt.
- Može da se nalazi u bilo kojem folderu Veb sajta.
- Daje instrukcije pretraživaču šta treba da pretražuje.
- Etika dobrog ponašanja pretraživača je da prati uputstva data u ovoj datoteci iako pretraživač ne mora da ih sledi.
- Veb sajt može da blokira određeni pretraživač ako se ne ponaša u skladu sa datim instrukcijama.

ROBOTS.TXT - PRIMER

User-agent: Google

Disallow:

User-agent: WebSpider

Disallow: /~joe/junk.html

Disallow: /~joe/foo.html

Disallow: /~joe/bar.html

User-agent: *

Disallow: /

OPTIMIZACIJA PRETRAŽIVANJA SA STRANE SAJTA

- Sitemap.xml – XML struktuirana datoteke, najčešće održavana od strane tehnologije koja pokreće Veb sajt.
- Nalazi se u folderu početne strane sajta.
- Sadrži spisak strana koje su od važnosti i njihove parametre.
- Parametri svake strane mogu biti URL putanja, datum poslednjeg ažuriranja, prioritet [0-1], frekvencija promene sadržaja.

SITEMAP.XML - PRIMER

```
<?xml version="1.0" encoding="UTF-8"?>
<loc>http://mywebiste.com/just-some-url.htm</loc>
<lastmod>2017-11-07T15:48:49+00:00</lastmod>
<changefreq>weekly</changefreq>
<priority>0.5</priority>
</url>
<url>
…
</url>
```

SITEMAP.XML I ROBOTS.TXT

- Mogu biti korisni ako se pravilno održavaju.
- Teže održavanja kod sajtova koji često menjaju sadržaj i strane.
- Istraživanja su pokazala da se ove dve datoteke ne održavaju pravilno.
- Većina sajtova ih i ne sadrži.
- Ako ih održava administrator umesto tehnologije koja pokreće Veb sajt, ove datoteke postaju još nepreciznije.
- Pretraživač po pravilu treba da prati uputstva iz robots.txt, dok sitemap.xml treba samo da konsultuje i uporedi sa svojim procenama.

KORISNI RADOVI I LINKOVI

- Cho, J. and H. Garcia-Molina - *The Evolution of the Web and Implications for an Incremental Crawler*
- L. Page, S. Brin et al - *The PageRank citation ranking: Bringing order to the web.*
- J Cho, H Garcia-Molina - *Effective page refresh policies for web crawlers*
- C Olston, M Najork - *Web crawling*
- Cai, R., et al. - *iRobot: an intelligent crawler for web forums*
- Jiang, J., et al. - *FoCUS: Learning to Crawl Web Forums*
- *Web Crawling - Microsoft Research - https://www.microsoft.com/en-us/research/publication/web-crawling/*
- Google patents pagerank - *https://www.google.com/patents/US6285999*

HVALA NA PAŽNJI

Email: milos_pavkovic@yahoo.com