Pretraga veba

Dragan Ivanović
dragan.ivanovic@uns.ac.rs

Katedra za informatiku, Fakultet tehničkih nauka, Novi Sad

2015.

Potreba

- Veb je danas nešto što se smatra normalnim načinom života
  - *"The Web has become the "new normal" in the American way of life; those who don't go online constitute an ever-shrinking minority." – [Pew Foundation report, January 2005]*
- Zašto kreirati veb sadržaj ako ga neće moći pronaći korisnici?
- Drugi vidovi pronalaženja nisu dobri zbog veličine veba
  - taksonomije
  - *bookmarks*
- Pretraga veba nam omogućava gotovo neograničenu selekciju onoga što želimo da saznamo ili kupimo
- Pomoću pretraživača veba moguća je agregacija interesa
  - kreiranje zajednica ljudi koji imaju isto interesovanje
  - online prodavnice koje prodaju usku paletu prodavnica
- Veb je danas ogromno, globalno tržište
- Zarada od reklama na veb pretraživačima je ogromno, jer na veb pretraživače dolazi jako puno ljudi

Veličina kolekcije

- Mnogo je veća i raste mnogo brže od kolekcija u ostalim IR sistemima
- Koliko je *host*-ova
- Koliko je stranica (statičkih)?
- Koja je količina podataka?
- NETCRAFT
- Broj stranica - numeričke procene
- Rast – usporen u odnosu na početno *"volume doubling every few months"*, ali i dalje veoma značajan

Način kreiranja sadržaja

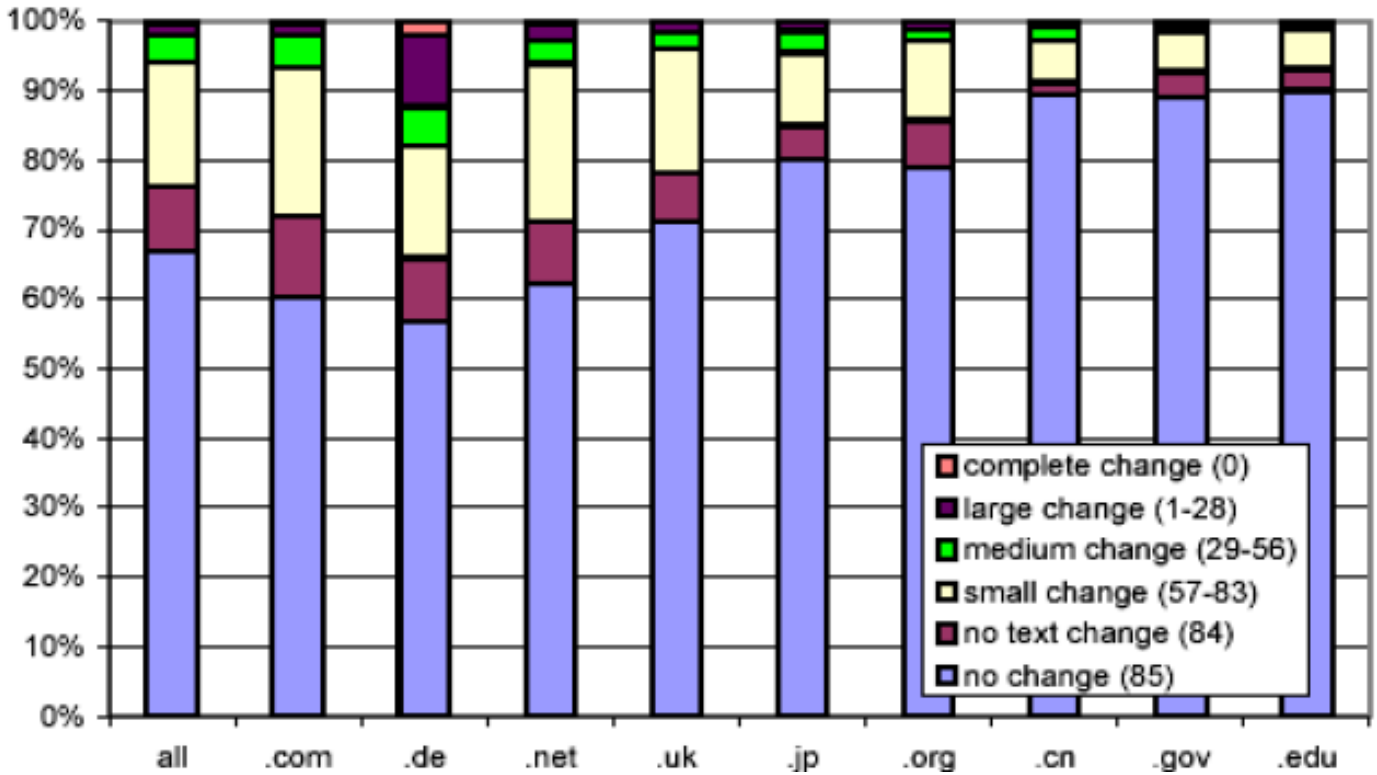
- Nema koordinacije u kreiranju sadržaja, demokratija u kreiranju sadržaja
- Distribuirano kreiranje sadržaja
- Povezivanje sadržaja
- Sadržaji uključuju istine, laži, zastarele informacije, kontradiktornosti
- Nestruktuirani (text, html, ...), polustruktuirani (XML, anotirane fotografije), struktuirani (baze podataka)

Način kreiranja sadržaja

- Dinamički sadržaji -Dinamički generisane html strane u momentu prijema zahteva
- HTTP zahtev obično ima karakter "?"
- Trenutno stanje leta AA129, raspoloživost soba u hotelu, prilikom generisanja html strane uzimaju se trenutni podaci iz baze podataka
- *www.yahoo.com/<anything>* je validna html strana
- *Spider*-i odnosno *Crawler*-i često ignorišu dinamičke sadržaje, da ne bi upali u maliciozne zamke
- Oprez: U nekim aplikacijama su vesti dinamički sadržaji, application-specific spidering
- *Statically indexable web* - sve što veb pretraživači indeksiraju

Česta izmena sadržaja

- Postoji više istraživanja, ali najveće je *Fetterly et al. (2003)*: Nekoliko puta su pregledani podaci o 150 miliona stranica tokom 11 nedelja istraživanja, postoji 85 različitih nivoa izmena sadržaja

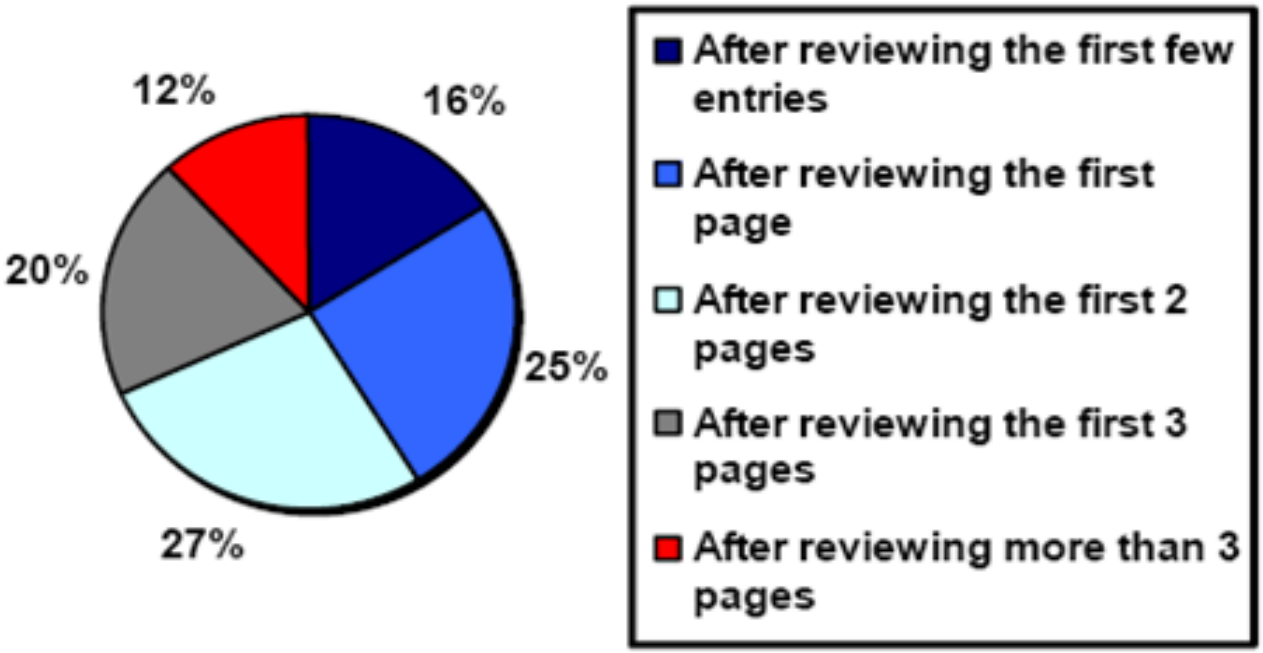


Tipični korisnici

- Ne gnjavi me operatorima i složenim upitnim konstrukcijama
- Kratki upiti - 80% upita ima jednu ili dve reči
- Neprecizni upiti
- Ne ulažu puno intelektualnih napora prilikom kreiranja upita
- 78% upita se dodatno ne reformuliše
- Velike razlike u potrebama, očekivanjima, znanju, tehničkoj i komunikacionoj opremi, godinama

Strpljenje korisnika

"When you perform a search on a search engine and don't find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)"



Potrebe korisnika

- Informacione - želi da nauči o nečemu, 2002 - 40%, 2004 - 65%
  - information retrieval
  - html
- Navigacione - želi da pronađe veb sajt neke kompanije, 2002 - 25%, 2004 - 15%
  - Wizzair
  - US open tennis
- Transakcione - želi da uradi nešto, 2002 - 35%, 2004 - 20%
  - preuzme datoteku - download lucene
  - pristupi servisu - weather novi sad
  - kupi nešto - buy Canon S410
- Gray areas
  - pronalaženje dobrog *hub*-a - car rental brasil
  - *explanatory search - see what's there*
  - često se ovo svrstava pod informacione potrebe

Korisnikova evaluacija veb pretraživača

- Prikaži mi rezultate korektno i brzo
- Jednostavan interfejs, tolerantan na moje greške, spell checking, did you mean, similar pages
- Ne davi me - pop ups, reklame, itd.
- Daj mi odgovore iz pouzdanih izvora, nemoj mi duplirati odgovore, daj mi različite odgovore, lepo ih organizuj
- Povrat važniji od preciznosti, nema veze što ima i onog što mi ne treba, samo da ima onog što mi treba
- Bitno je da je to što mi treba među prvim rezultatima, bitna je preciznost prvih k odgovora, 85% gleda samo prvu stranu sa rezultatima
- Mišljenje jednog korisnika može biti beznačajno, ali mišljenje velike količine korisnika se mora analizirati i uzeti u obzir - log mining

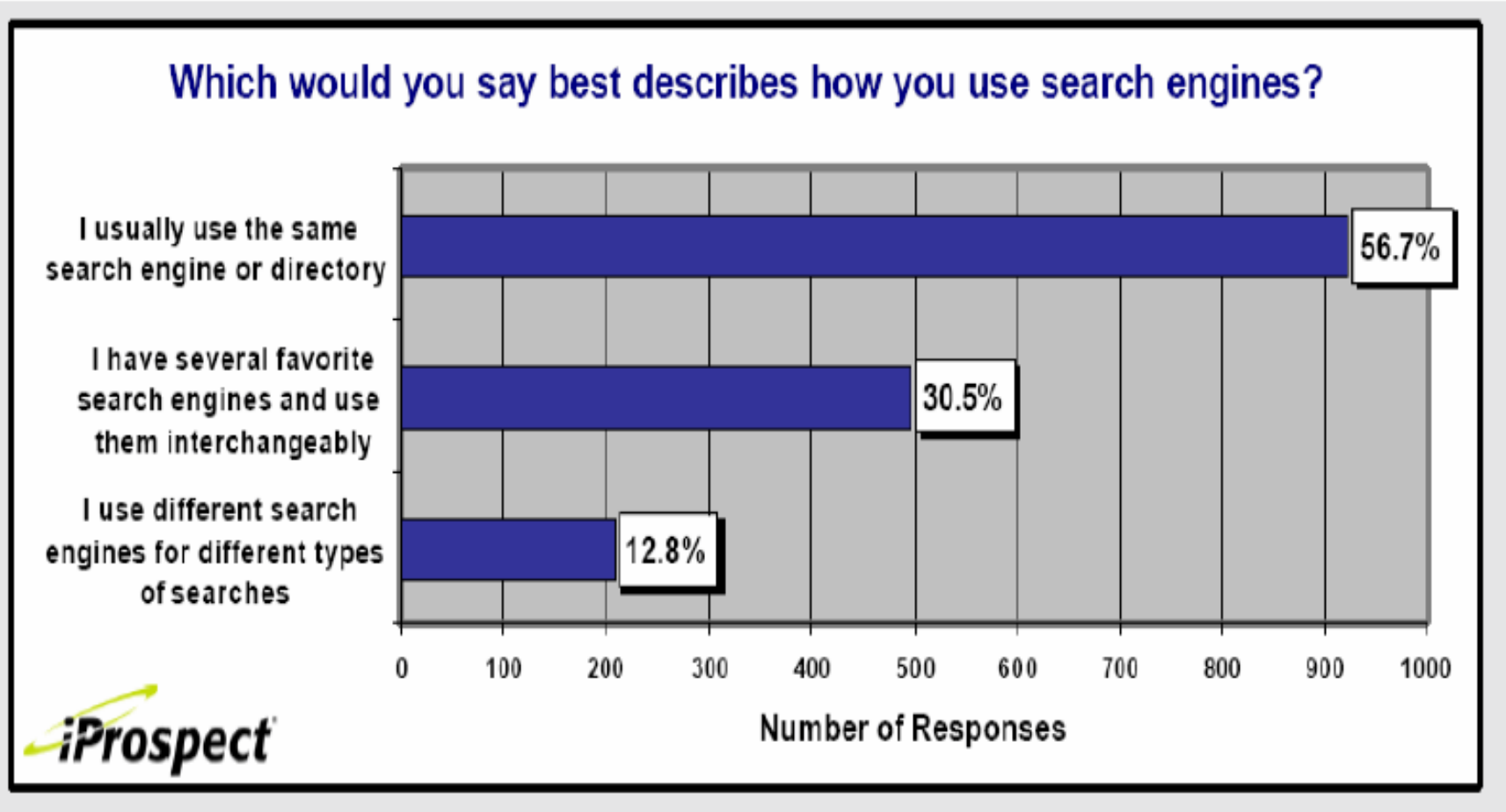
The need behind the query

- Podsećanje, rezultati u IR treba da budu u skladu sa informacionom potrebom, a ne upitom
- Čitanje između redova
- (Realni) korisnici su zadovoljni ako nakaradno postave upit, jer ni sami nisu sigurni šta traže, a veb pretraživač ih uputi na ono što im treba
- Ne odnosi se samo na *did you mean*
- Koriste se i tehnike *data mining*-a
- Reformulacija prirodnih jezika tako da se dobije bolja lista odgovora, ne zaboravite i da slabo pismeni koriste veb pretraživače
- Ako upit ima samo ime grada najpre će biti rezultati sa mapom i turističkim vodičima za taj grad

The need behind the query

- Personalizacija pretrage - na *Google* možete videti i svoju statistiku pretraga, a onda *Google* to koristi za Vaše buduće pretrage, slično je i kod ostalih veb pretraživača
- *Google.rs* i *Google.fr* mogu dati drugačije rezultate za isti upit
- Čak i da nemate nalog veb pretraživača ili da niste ulogovani, pomoću korisničke sesije se mogu utvrditi Vaši prethodni upiti i na osnovu njih izvesti zaključci
- Nekad se i na osnovu IP adrese utvrđuje odakle ste
- Restrikcija rezultata - uklanjanje neodgovarajućih
- Izmena rangiranja, prvo generičko rangiranje, a onda personalizuj rangiranje prvih n rezultata

Lojalnost korisnika



Relevantnost dokumenata

- Milijarde stranica koje predstavljaju spam
- Za rangiranje rezultata kod veb pretraživača se koristi mnogo više parametara, ne samo tf-idf
- Cilj je maksimalno razdvojili kvalitetne i nekvalitetne veb sadržaje - veb sadržaji uključuju istine, laži, zastarele informacije, kontradiktornosti
- Veb sadržaji su povezani velikom količinom linkova - u proseku postoji preko 8 linkova sa jedne strane, kompleksan graf veba
- Iz ovih grafova se mogu crpeti informacije o kvalitetu veb sadržaja - **analiza linkova** je veoma značajna za pretragu veba

Približni duplikati

- Veb je prepun duplih sadržaja
- Potpuni duplikati se mogu detektovati ali njih nema toliko puno
- Ali postoji mnogo, mnogo slučajeva približnih duplikata (Near duplicates) - 35-40%
  - Sintaksno 35-40% (približnih) duplikata (*Broder et al., 1997*) - Različit je samo datum poslednje izmene, ili je različito zaglavlje, na različitim forumima isto pitanje i isti odgovori, itd.
  - Semantički - ne zna se, ali sigurno je značajan procenat

Detekcija potpunih duplikata

- Još jednom, ovih duplikata **nema** mnogo
- Potpuni duplikati se mogu detektovati *fingerprinting* algoritmima
  - Veliki niz podataka se pretvori u jedinstveni niz bitova koji ga identifikuje, na primer hash funkcija
  - Isto kao što čoveka identifikuje otisak prstiju
  - wiki stranica - link

Detekcija približnih duplikata

- Još jednom, ovih duplikata **ima** mnogo
- Potrebna je mera sličnosti - Edit distance
- Mora se odrediti neki treshold - sličnost >80% se uzima za približni duplikat
- Mera sličnosti nije tranzitivna, ali je u nekim situacijama veb pretraživači koriste tranzitivno

Detekcija približnih duplikata

- Dokument se izdeli da delove
- ili se izdeli na *shingles* (N-grame reči)
- Formiramo skup istih delova ili *shingles* dva dokumenta koji se porede - presek
- Formiramo i uniju delova ili *shingles* dva dokumenta koji se porede
- Količnik veličine preseka i unije je mera sličnosti
- Ovo računanje je nekad skupo, pa se uzima samo uzorak delova ili *shingles* iz dokumenata - *Sketch*
- Postoje tehnike za odabir uzoraka - uvek pitanje da li je uzorak dobar visi u vazduhu
- Matrice i *Jaccard*-ov proizvod se takođe koriste za brže računanje mere sličnosti

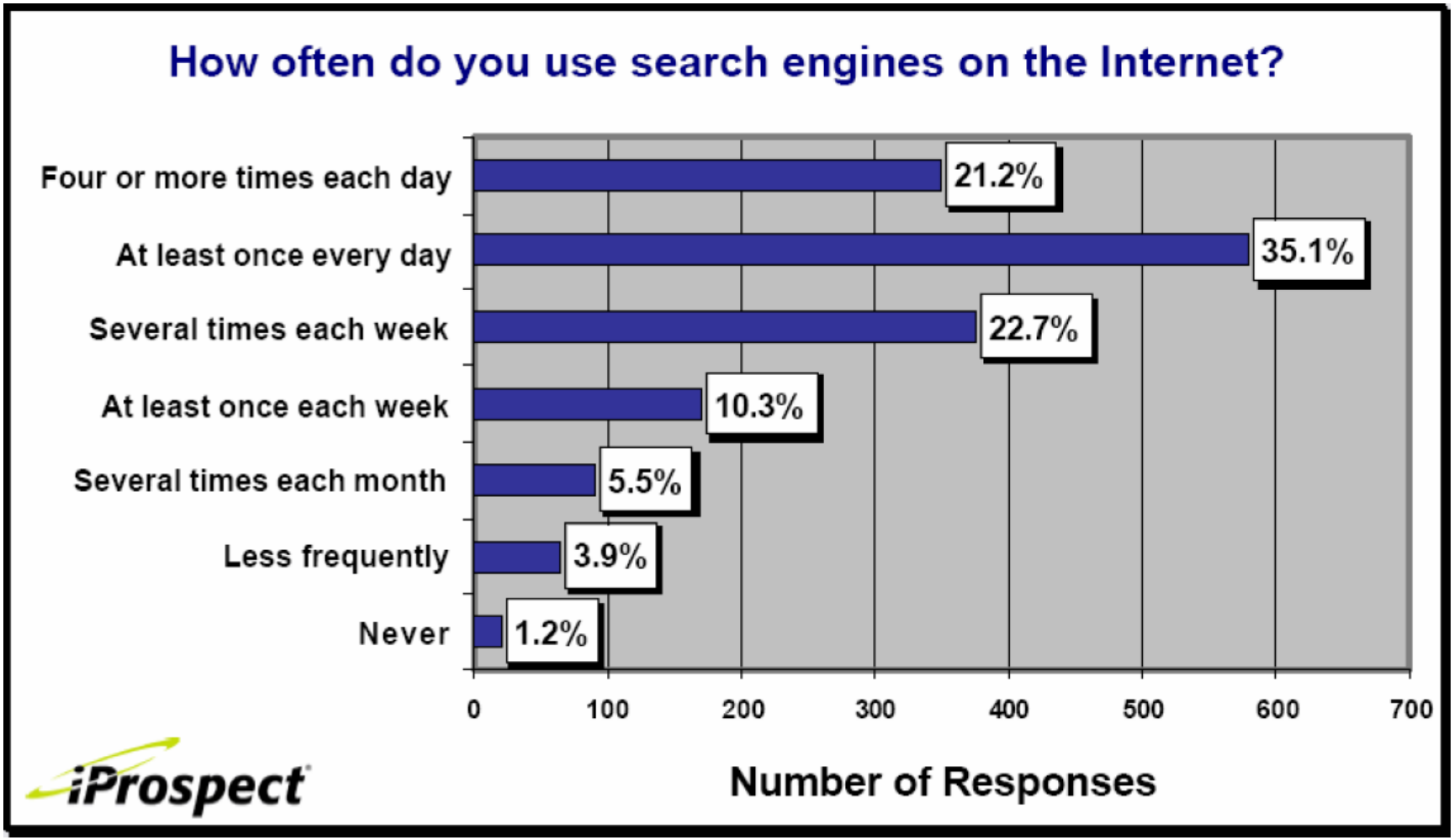
Istorija

- Prvi veb pretraživači bazirani na ključnim rečima su se pojavili u periodu 1995-1997
  - *Altavista*, *Excite*, *Infoseek*, *Inktomi*, *Lycos*
- *Paid search rangiranje*
  - *Goto* (kasnije *Overture.com*, *Yahoo* 2003. godine ih je kupio za 1.63 milijardi dolara)
  - Rangiranje određenog sajta na upit je zavisilo od toga koliko su platili
  - Aukcija za ključne reči: _casino je bio jako skup

Istorija

- 1998+ - Link bazirano rangiranje uvedeno od strane *Google*-a
  - Oduvali su konkurenciji, korisnici su bili zadovoljni
  - U to vreme *Goto/Overture* je imao godišnji prihod od preko milijardu dolara
- *Google* kasnije dodao sekciju *Ads* koja je nezavisna od osnovnih rezultata pretrage (desna polovina ekrana) - slično usvojio i Yahoo!
- *Google* danas ima najviše korisnika, samim tim verovatno i najviše zarade

Koliko često?



Ko je ko

comScore Explicit Core Search Share Report*
February 2014 vs. January 2014
Total U.S. – Home & Work Locations
Source: comScore qSearch

Core Search Entity	Explicit Core Search Share (%)		
	Jan-14	Feb-14	Point Change
Total Explicit Core Search	*100.0%*	*100.0%*	*N/A*
Google Sites	67.6%	67.5%	-0.1
Microsoft Sites	18.3%	18.4%	0.1
Yahoo Sites	10.4%	10.3%	-0.1
Ask Network	2.4%	2.4%	0.0
AOL, Inc.	1.3%	1.3%	0.0

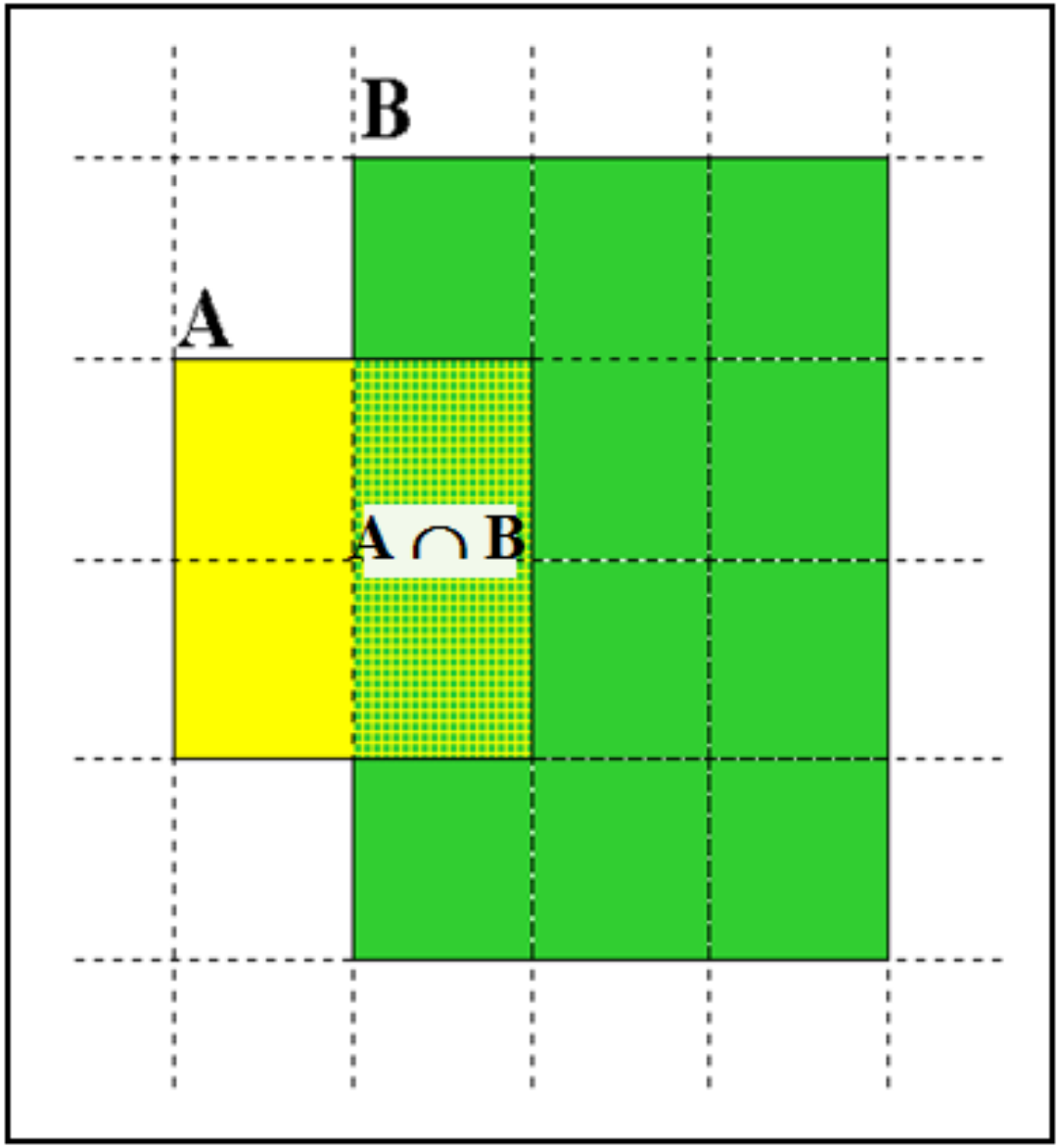
Veličina kolekcije veb pretraživača

- Kolika je kolekcija koju indeksira neki veb pretraživač
- Pojam broja strana koje indeksira neki veb pretraživač je relativno dobro definisan, ali postoje i neki problemi
  - Pretraživači indeksiraju i stranice koje još nisu preuzeli putem *crawler*-a (*anchortext*)
  - Restrikcije: indeksirano je samo prvih n reči u stranici
- Relativno poređenje pokrivanja veb pretraživača

Razlike pokrivenosti veb pretraživača

- Različiti parametri veb pretraživača - *max url depth, max count/host, anti-spam rules, priority rules*,itd.
- Za isti URL se indeksiraju različiti sadržaji - *frames, meta-keywords, document restrictions, document extensions*, itd.

Poređenje pokrivenosti veb pretraživača



Poređenje pokrivenosti veb pretraživača

- Uzorci - random *URL*-ovi iz pretraživača A
- Provera - da li tako odabrani *URL* postoji i u B
- Uraditi i obrnuto
- Odnos veličine kolekcije koju pokriva pretraživač A i veličine kolekcije koju pokriva pretraživač B je srazmeran odnosu broja slučajnih uzoraka iz B pronadjenih u A i broja slučajnih uzoraka iz A pronadjenih u B
- Kako birati uzorke
- Kako vršiti provere

Odabir uzoraka

- Generisanje random upita
  - *Lexicon*: 400,000+ reči prikupljenih putem crawler-a
  - Odabere se nekoliko termina i izvrši konjukcija: w1 and w2
- Uzmu se *URL*-ovi prvih 100 rezultata kada se ovaj upit izvrši na pretraživaču A
- Uzme se slučajan predstavnik od ovih 100 i proveri se njegova prisutnost u pretraživaču B
- Postoje i drugi mehanizmi za odabir uzoraka: *random search, random IP addresses, random walks*

Provera

- Da li pretraživač B indeksira dokument D
- Preuzmi D, uzmi njegovu listu reči
- Napravi konjukciju između 8 reči male frekvencije i postavi taj upit pretraživaču B
- Proveri da li je među odgovorima D
- Problemi:
  - Približni duplikati
  - Redirekcije
  - Da li je 8 reči dovoljno? Da li je previše odgovora?