Introduction to

Information Retrieval

ΠΛΕ70: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Διάλεξη 10: Βασικά Θέματα Αναζήτησης στον Παγκόσμιο Ιστό.

Τι θα δούμε σήμερα;

- Τι ψάχνουν οι χρήστες
- Διαφημίσεις
- Spam
- Πόσο μεγάλος είναι ο Ιστός;

ΟΙ ΧΡΗΣΤΕΣ

Ανάγκες Χρηστών

- Ποιοι είναι οι χρήστες;
- Μέσος αριθμός λέξεων ανά αναζήτηση 2-3
- Σπάνια χρησιμοποιούν τελεστές

Ανάγκες Χρηστών

Need [Brod02, RL04]

- **Informational** (πληροφοριακά ερωτήματα) – θέλουν να μάθουν (learn) για κάτι (~40% / 65%)
  - Συνήθως, όχι μια μοναδική ιστοσελίδα, συνδυασμός πληροφορίας από πολλές ιστοσελίδες

`Low hemoglobin`

- **Navigational** (ερωτήματα πλοήγησης) – θέλουν να πάνε (go) σε μια συγκεκριμένη ιστοσελίδα (~25% / 15%)
  - Μια μοναδική ιστοσελίδα, το καλύτερο μέτρο = ακρίβεια στο 1 (δεν ενδιαφέρονται γενικά για ιστοσελίδες που περιέχουν τους όρους United Airlines)

`United Airlines`

Ανάγκες Χρηστών

Transactional (ερωτήματα συναλλαγής) – θέλουν να κάνουν (do) κάτι (σχετιζόμενο με το web) (~35% / 20%)

- Προσπελάσουν μια υπηρεσία (Access a service)
- Να κατεβάσουν ένα αρχείο (Downloads)
- Να αγοράσουν κάτι
- Να κάνουν κράτηση

`Seattle weather`

`Mars surface images`

`Canon S410`

- **Γκρι περιοχές** (Gray areas)
  - Find a good hub
  - Exploratory search "see what's there"

`Car rental Brasil`

Τι ψάχνουν;

Δημοφιλή ερωτήματα

- http://www.google.com/trends/hottrends

Και ανά χώρα

Τα ερωτήματα ακολουθούν επίσης power law κατανομή

Ανάγκες Χρηστών

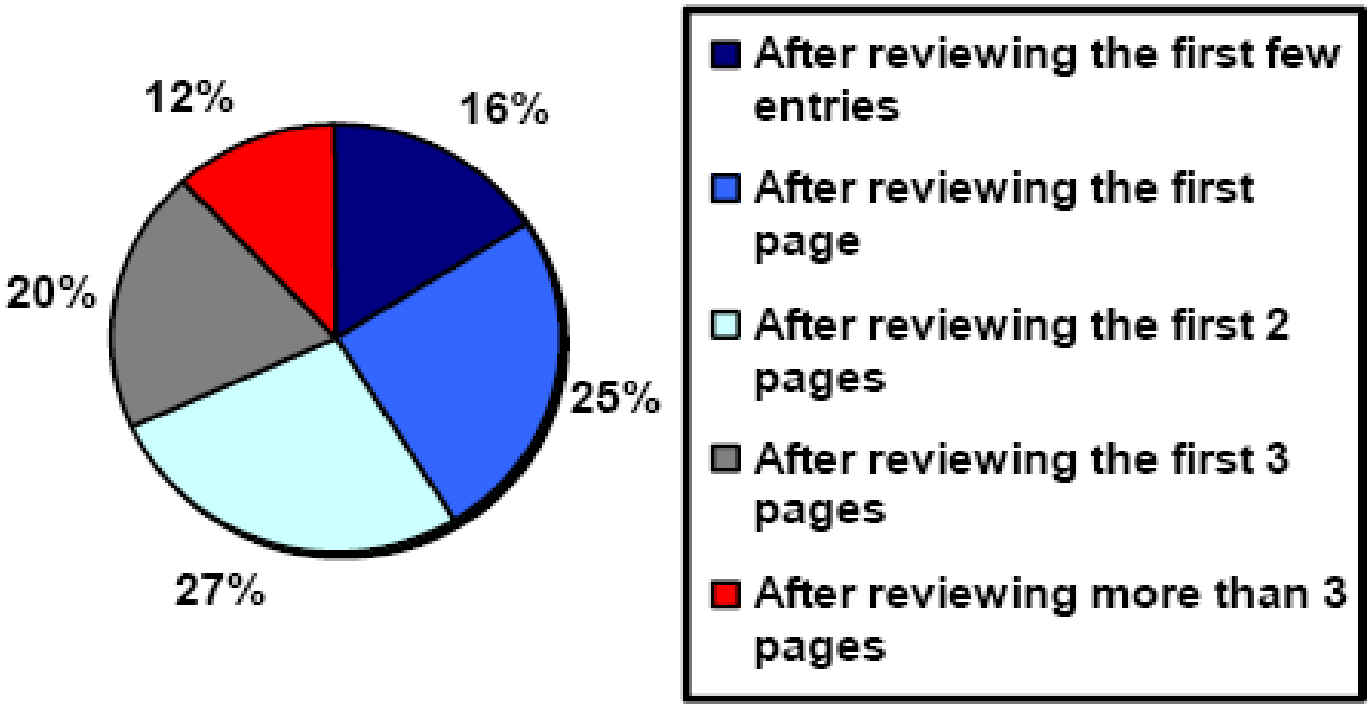
Επηρεάζει (ανάμεσα σε άλλα)

- την καταλληλότητα του ερωτήματος για την παρουσίαση *διαφημίσεων*

- τον *αλγόριθμο/αξιολόγηση*, για παράδειγμα για ερωτήματα πλοήγησης ένα αποτέλεσμα ίσως αρκεί, για τα άλλα (και κυρίως πληροφοριακά) ενδιαφερόμαστε για την περιεκτικότητα/ανάκληση

Πόσα αποτελέσματα βλέπουν οι χρήστες

"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)"



(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)

Πως μπορούμε να καταλάβουμε τις προθέσεις (intent) του χρήστη;

Guess user intent *independent of context*:

- Spell correction
- Precomputed "typing" of queries

Better: Guess user intent *based on context*:

- Geographic context (slide after next)
- Context of user in this session (e.g., previous query)
- Context provided by personal profile (Yahoo/MSN do this, Google claims it doesn't)

Examples of Typing Queries

Calculation: 5+4
Unit conversion: 1 kg in pounds
Currency conversion: 1 euro in kronor
Tracking number: 8167 2278 6764
Flight info: LH 454
Area code: 650
Map: columbus oh
Stock price: msft
Albums/movies etc: coldplay

Geographical Context

Three relevant locations

1. Server (nytimes.com → New York)
2. Web page (nytimes.com article about Albania)
3. User (located in Palo Alto)

Locating the user

- IP address
- Information provided by user (e.g., in user profile)
- Mobile phone

Geo-tagging: Parse text and identify the coordinates of the geographic entities

Example: East Palo Alto CA → Latitude: 37.47 N, Longitude: 122.14 W

✓ Important NLP problem

Geographical Context

How to use context to modify query results:

- Result restriction: Don't consider inappropriate results
  - For user on google.fr only show .fr results
- Ranking modulation: use a rough generic ranking, rerank based on personal context

Contextualization / personalization is an area of search with a lot of potential for improvement.

Αξιολόγηση από τους χρήστες

- Relevance and validity of results
  - Precision at 1? Precision above the fold?
  - Comprehensiveness – must be able to deal with obscure queries
    - Recall matters when the number of matches is very small
- UI (User Interface) – Simple, no clutter, error tolerant
  - No annoyances: pop-ups, etc.
- Trust – Results are objective
- Coverage of topics for polysemic queries
  - Diversity, duplicate elimination

Αξιολόγηση από τους χρήστες

- Pre/Post process tools provided
  - Mitigate user errors (auto spell check, search assist,…)
  - Explicit: Search within results, more like this, refine ...
  - Anticipative: related searches
- Deal with idiosyncrasies
  - Web specific vocabulary
    - Impact on stemming, spell-check, etc.
  - Web addresses typed in the search box

ΔΙΑΦΗΜΙΣΕΙΣ

Ads

Graphical graph banners on popular web sites (branding)

- ***cost per mil (CPM) model***: the cost of having its banner advertisement displayed 1000 times (also known as impressions)
- ***cost per click (CPC) model***: number of clicks on the advertisement (leads to a web page set up to make a purchase)

✓ brand promotion vs transaction-oriented advertising

Brief (non-technical) history

- Early keyword-based engines ca. 1995-1997
  - Altavista, Excite, Infoseek, Inktomi, Lycos
- Paid search ranking: Goto (morphed into Overture.com → Yahoo!)
  - Your search ranking depended on how much you paid
  - Auction for keywords: ***casino*** was expensive!

Ads in Goto

In response to the query *q*, Goto

- return the pages of all advertisers who bid for *q*, ordered by their bids.
- when the user clicked on one of the returned results, the corresponding advertiser payment to Goto
  - Initially, payment equal to bid for q
  - *Sponsored search* or *Search advertising*

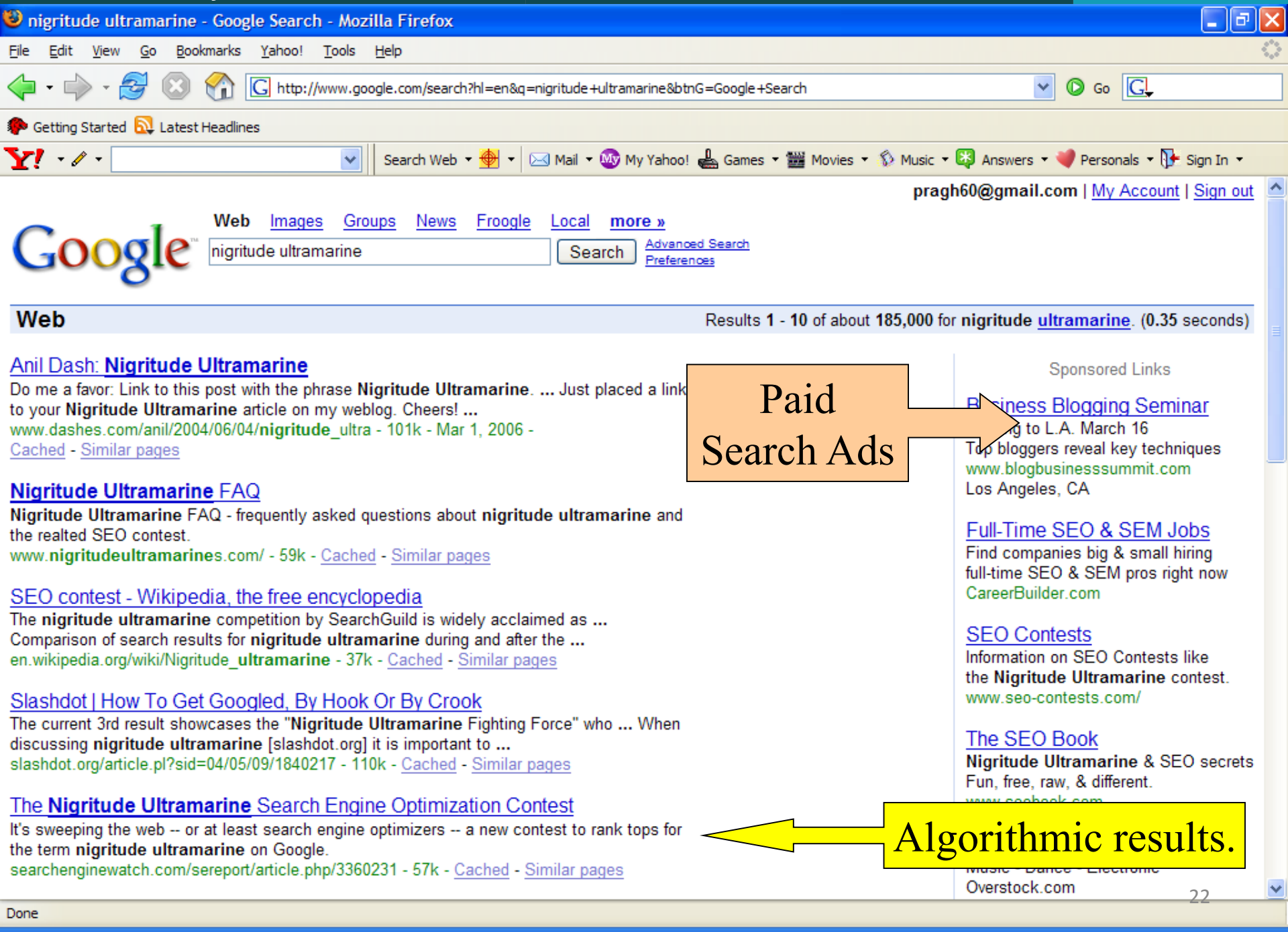
Ads in Goto


Wilmington real estate
Access 75% of all users now!
Premium Listings reach 75% of all Internet users. Sign up for Premium Listings today!
1. Wilmington Real Estate - Buddy Blake
Wilmington's information and real estate guide. This is your on
anything to do with Wilmington.
www.buddyblake.com (Cost to advertiser: $0.28)
2. Coldwell Banker Sea Coast Realty
Wilmington's number one real estate company.
www.cbseacoast.com (Cost to advertiser: $0.37)
3. Wilmington, NC Real Estate Becky Bullard
Everything you need to know about buying or selling a home c
on my Web site!
www.iwwc.net (Cost to advertiser: $0.25)

Ads

Provide

- ***pure search results*** (generally known as algorithmic or organic search results) as the primary response to a user's search,
- together with ***sponsored search results*** displayed separately and distinctively to the right of the algorithmic results.


nigritude ultramarine - Google Search - Mozilla Firefox
File Edit View Go Bookmarks Yahoo! Tools Help
http://www.google.com/search?hl=en&q=nigritude+ultramarine&btnG=Google+Search
Getting Started Latest Headlines
Search Web Mail My Yahoo! Games Movies Music Answers Personals Sign In
pragh60@gmail.com | My Account | Sign out
Google
Web Images Groups News Froogle Local more »
nigritude ultramarine
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 185,000 for nigritude ultramarine. (0.35 seconds)
Anil Dash: Nigritude Ultramarine
Do me a favor: Link to this post with the phrase Nigritude Ultramarine. ... Just placed a link to your Nigritude Ultramarine article on my weblog. Cheers! ...
www.dashes.com/anil/2004/06/04/nigritude_ultra - 101k - Mar 1, 2006 -
Cached - Similar pages
Nigritude Ultramarine FAQ
Nigritude Ultramarine FAQ - frequently asked questions about nigritude ultramarine and the realted SEO contest.
www.nigritudeultramarines.com/ - 59k - Cached - Similar pages
SEO contest - Wikipedia, the free encyclopedia
The nigritude ultramarine competition by SearchGuild is widely acclaimed as ...
Comparison of search results for nigritude ultramarine during and after the ...
en.wikipedia.org/wiki/Nigritude_ultramarine - 37k - Cached - Similar pages
Slashdot | How To Get Googled, By Hook Or By Crook
The current 3rd result showcases the "Nigritude Ultramarine Fighting Force" who ... When discussing nigritude ultramarine [slashdot.org] it is important to ...
slashdot.org/article.pl?sid=04/05/09/1840217 - 110k - Cached - Similar pages
The Nigritude Ultramarine Search Engine Optimization Contest
It's sweeping the web -- or at least search engine optimizers -- a new contest to rank tops for the term nigritude ultramarine on Google.
searchenginewatch.com/sereport/article.php/3360231 - 57k - Cached - Similar pages
Paid Search Ads
Sponsored Links
Business Blogging Seminar
g 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 secrets
Fun, free, raw, & different.
Overstock.com
Algorithmic results.
Done

Ads

- **Search Engine Marketing (SEM)**

Understanding how search engines do ranking and how to allocate marketing campaign budgets to different keywords and to different sponsored search engines

- **Click spam**: clicks on sponsored search results that are not from bona fide search users.
  - For instance, a devious advertiser

Ads

Paid inclusion: pay to have one's web page included in the search engine's index

Different search engines have *different policies* on whether to allow paid inclusion, and whether such a payment has any effect on ranking in search results.

Similar problems with TV/newspapers

How are ads ranked?

- Advertisers *bid for keywords* – sale by auction.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on their ad.
- Important area for search engines – *computational advertising.*
  - an additional fraction of a cent from each ad means billions of additional revenue for the search engine.

How are ads ranked?

- How does the auction determine an ad's rank and the price paid for the ad?
  - Basis is a second price auction

Google's second price auction

advertiser	bid	CTR	ad rank	rank	paid
A	$4.00	0.01	0.04	4	(minimum)
B	$3.00	0.03	0.09	2	$2.68
C	$2.00	0.06	0.12	1	$1.51
D	$1.00	0.08	0.08	3	$0.51

- bid: maximum bid for a click by advertiser
- CTR: click-through rate: when an ad is displayed, what percentage of time do users click on it? CTR is a measure of relevance.
- ad rank: bid × CTR: this trades off (i) how much money the advertiser is willing to pay against (ii) how relevant the ad is
- rank: rank in auction
- paid: second price auction price paid by advertiser

Google's second price auction

advertiser	bid	CTR	ad rank	rank	paid
A	\$4.00	0.01	0.04	4	(minimum)
B	\$3.00	0.03	0.09	2	\$2.68
C	\$2.00	0.06	0.12	1	\$1.51
D	\$1.00	0.08	0.08	3	\$0.51

Second price auction: The advertiser pays the minimum amount necessary to maintain their position in the auction (plus 1 cent).

$price_1 \times CTR_1 = bid_2 \times CTR_2$ (this will result in $rank_1 = rank_2$)

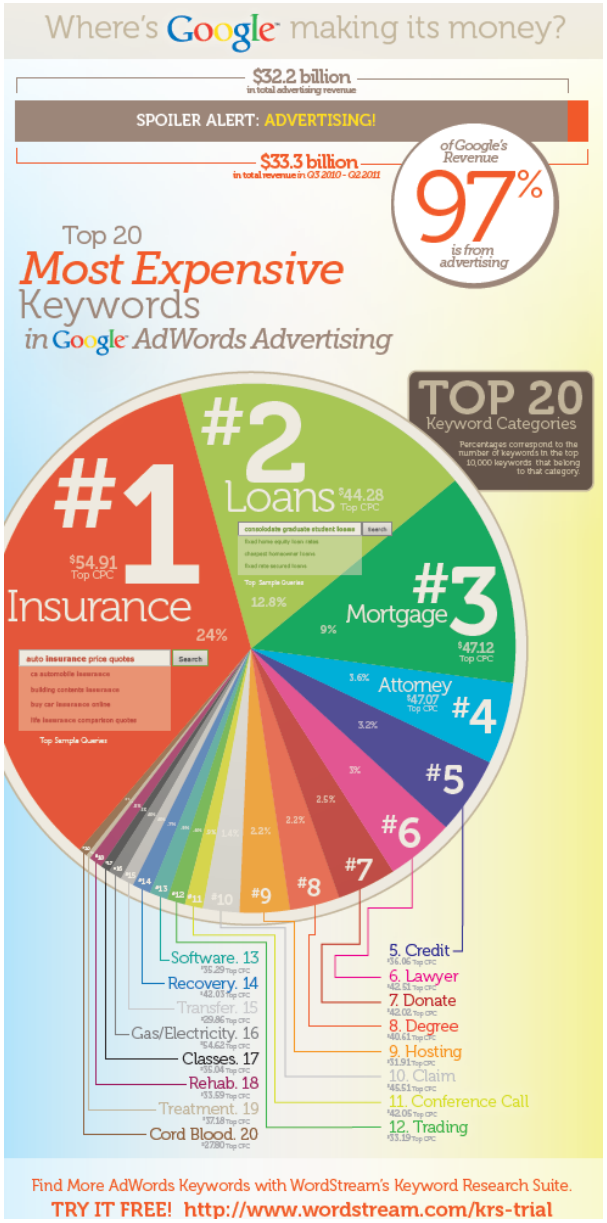
$price_1 = bid_2 \times CTR_2 / CTR_1$

$p_1 = bid_2 \times CTR_2/CTR_1 = 3.00 \times 0.03/0.06 = 1.50$
$p_2 = bid_3 \times CTR_3/CTR_2 = 1.00 \times 0.08/0.03 = 2.67$
$p_3 = bid_4 \times CTR_4/CTR_3 = 4.00 \times 0.01/0.08 = 0.50$

Keywords with high bids

According to http://www.cwire.org/highest-paying-search-terms/

$69.1	mesothelioma treatment options
$65.9	personal injury lawyer michigan
$62.6	student loans consolidation
$61.4	car accident attorney los angeles
$59.4	online car insurance quotes
$59.4	arizona dui lawyer
$46.4	asbestos cancer
$40.1	home equity line of credit
$39.8	life insurance quotes
$39.2	refinancing
$38.7	equity line of credit
$38.0	lasik eye surgery new york city
$37.0	2nd mortgage
$35.9	free car insurance quote


Where's Google making its money?
$32.2 billion
in total advertising revenue
SPOILER ALERT: ADVERTISING!
$33.3 billion
in total revenue in Q3 2010 - Q2 2011
of Google's Revenue
97%
is from advertising
Top 20
Most Expensive
Keywords
in Google AdWords Advertising
TOP 20
Keyword Categories
Percentages correspond to the number of keywords in the top 10,000 keywords that belong to that category.
#1
Insurance
$54.91
Top CPC
24%
#2
Loans
$44.28
Top CPC
12.8%
#3
Mortgage
$47.12
Top CPC
9%
#4
Attorney
$47.07
Top CPC
3.6%
#5
3.2%
#6
3%
#7
2.6%
#8
2.2%
#9
2.2%
#10
1.4%
#11
#12
#13
#14
#15
#16
#17
#18
#19
#20
5. Credit
6. Lawyer
7. Donate
8. Degree
9. Hosting
10. Claim
11. Conference Call
12. Trading
Software. 13
Recovery. 14
Transfer. 15
Gas/Electricity. 16
Classes. 17
Rehab. 18
Treatment. 19
Cord Blood. 20
Find More AdWords Keywords with WordStream's Keyword Research Suite.
TRY IT FREE! http://www.wordstream.com/krs-trial

Search ads: A win-win-win?

- The search engine company gets revenue every time somebody clicks on an ad.
- The user only clicks on an ad if they are interested in the ad.
  - Search engines punish misleading and nonrelevant ads.
  - As a result, users are often satisfied with what they find after clicking on an ad.
- The advertiser finds new customers in a cost-effective way.

Not a win-win-win: Keyword arbitrage

- Buy a keyword on Google
- Then redirect traffic to a third party that is paying much more than you are paying Google.
  - E.g., redirect to a page full of ads
- This rarely makes sense for the user.
- Ad spammers keep inventing new tricks.
- The search engines need time to catch up with them.

32

Not a win-win-win: Violation of trademarks

- Example: geico
- During part of 2005: The search term "geico" on Google was bought by competitors.
- Geico lost this case in the United States.
- Louis Vuitton lost similar case in Europe (2010).

- It's potentially misleading to users to trigger an ad off of a trademark if the user can't buy the product on the site.

SPAM
(SEARCH ENGINE OPTIMIZATION)

The trouble with paid search ads

- It costs money. What's the alternative?

Search Engine Optimization (SEO):

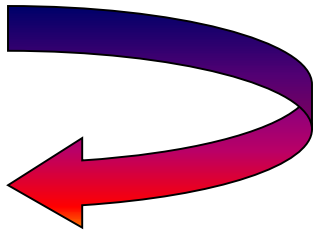
  - "Tuning" your web page to rank highly in the algorithmic search results for select keywords
  - Alternative to paying for placement
  - Thus, intrinsically a marketing function

- Performed by companies, webmasters and consultants ("Search engine optimizers") for their clients
- Some perfectly legitimate, some very shady

Η απλούστερη μορφή

- Οι μηχανές πρώτης γενιάς βασίζονταν πολύ στο *tf/idf*
  - Οι πρώτες στην κατάταξη ιστοσελίδας για το ερώτημα **`maui resort`** ήταν αυτές που περιείχαν τα περισσότερα **`maui`** και **`resort`**
- SEOs απάντησαν με πυκνή επανάληψη των επιλεγμένων όρων
  - π.χ., **`maui resort maui resort maui resort`**
  - Συχνά, οι επαναλήψεις στο ίδιο χρώμα με background της ιστοσελίδα
    - Οι επαναλαμβανόμενοι όροι έμπαιναν στο ευρετήριο από crawlers
    - Αλλά δεν ήταν ορατοί από τους ανθρώπους στους browsers

Απλή πυκνότητα όρων δεν είναι αξιόπιστο ΑΠ σήμα



Παραλλαγές «keyword stuffing»

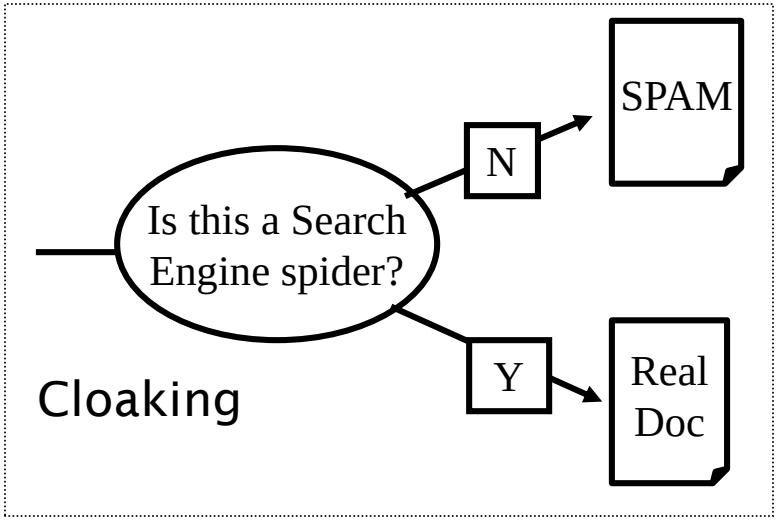
a web page loaded with keywords in the meta tags
or in content of a web page (outdated)

- Παραπλανητικά meta-tags, υπερβολική επανάληψη
- Hidden text with colors, position text behind the image, style sheet tricks, etc.

Meta-Tags =
"… London hotels, hotel, holiday inn, hilton, discount, booking, reservation, sex, mp3, britney spears, viagra, …"

Cloaking (Απόκρυψη)

- Παρέχει διαφορετικό περιεχόμενο ανάλογα αν είναι ο μηχανισμός σταχυολόγησης (search engine spider) ή ο browser κάποιου χρήστη
- DNS cloaking: Switch IP address. Impersonate



Άλλες τεχνικές παραπλάνησης (spam)

- Doorway pages
  - Pages optimized for a single keyword that re-direct to the real target page
  - If a visitor clicks through to a typical doorway page from a search engine results page, redirected with a fast *Meta refresh* command to another page.

- Lander page:

optimized for a single keyword or a misspelled domain name, designed to attract surfers who will then click on ads

Άλλες τεχνικές παραπλάνησης (spam)

- Link spamming
  - Mutual admiration societies, hidden links, awards
  - *Domain flooding:* numerous domains that point or re-direct to a target page
  - Pay somebody to put your link on their highly ranked page
  - Leave comments that include the link on blogs
- Robots (bots)
  - Fake query stream – rank checking programs
    - "Curve-fit" ranking programs of search engines
  - Millions of submissions via Add-Url

The war against spam

- Quality signals - Prefer authoritative pages based on:
  - Votes from authors (linkage signals)
  - Votes from users (usage signals)
- Policing of URL submissions
  - Anti robot test
- Limits on meta-keywords
- Robust link analysis
  - Ignore statistically implausible linkage (or text)
  - Use link analysis to detect spammers (guilt by association)
- Spam recognition by machine learning
  - Training set based on known spam
- Family friendly filters
  - Linguistic analysis, general classification techniques, etc.
  - For images: flesh tone detectors, source text analysis, etc.
- Editorial intervention
  - Blacklists
  - Top queries audited
  - Complaints addressed
  - Suspect pattern detection

More on spam

- Web search engines have policies on SEO practices they tolerate/block
  - http://help.yahoo.com/help/us/ysearch/index.html
  - http://www.google.com/intl/en/webmasters/
- Adversarial IR (Ανταγωνιστική ανάκτηση πληροφορίας): the unending (technical) battle between SEO's and web search engines
- Research http://airweb.cse.lehigh.edu/

Check out: Webmaster Tools (Google)

SIZE OF THE WEB

Ποιο είναι το μέγεθος του web ?

- Θέματα
  - Στην πραγματικότητα, ο web είναι άπειρος
    - Dynamic content, e.g., calendars
    - Soft 404: www.yahoo.com/<anything> is a valid page
  - Static web contains syntactic duplication, mostly due to mirroring (~30%)
  - Some servers are seldom connected
- Ποιο νοιάζει;
  - Media, and consequently the user
  - Σχεδιαστές μηχανών
  - Την πολιτική crawl - αντίκτυπο στην ανάκληση.

Τι μπορούμε να μετρήσουμε;

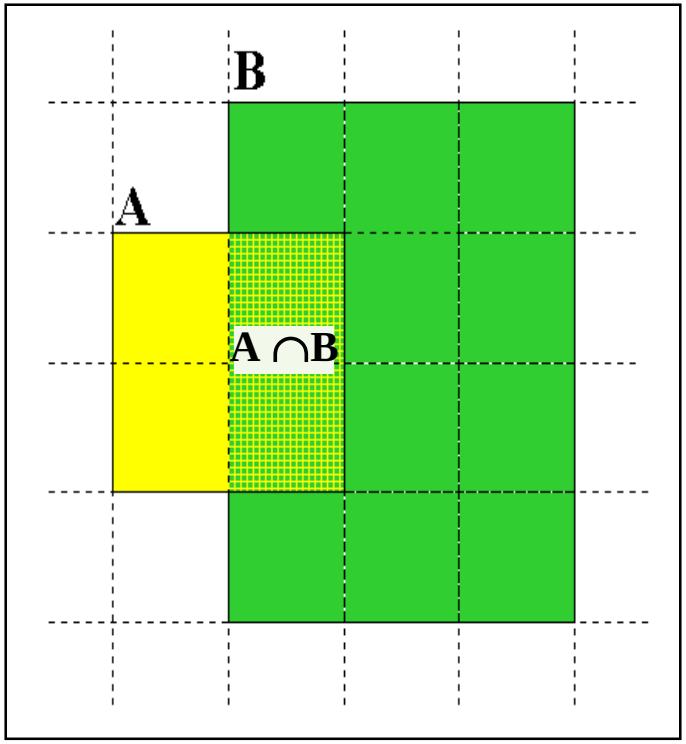
Το σχετικό μέγεθος των μηχανών αναζήτησης

- The notion of a page being indexed is still *reasonably* well defined.
- Already there are problems
  - Document extension: e.g., engines index pages not yet crawled, by indexing anchortext.
  - Document restriction: All engines restrict what is indexed (first *n* words, only relevant words, etc.)
  - Multi-tier indexes (access only top-levels)

New definition?

- The statically indexable web is whatever search engines index.
  - IQ is whatever the IQ tests measure.
- Different engines have different preferences
  - max url depth, max count/host, anti-spam rules, priority rules, etc.
- Different engines index different things under the same URL:
  - frames, meta-keywords, document restrictions, document extensions, ...

Μέγεθος μηχανών αναζήτησης



Relative Size from Overlap
Given two engines A and B

1. **Sample** URLs randomly from A
2. **Check** if contained in B and vice versa

$$A \cap B = (1/2) * \text{Size } A$$

$$A \cap B = (1/6) * \text{Size } B$$

$$(1/2)*\text{Size } A = (1/6)*\text{Size } B$$

$$\therefore \text{ Size } A / \text{ Size } B = (1/6)/(1/2) = 1/3$$

Each test involves: (i) Sampling (ii) Checking

Δειγματοληψία (Sampling) URLs

Ideal strategy: Generate a *random URL*

- Problem: Random URLs are hard to find (and sampling distribution should reflect "user interest")
- Approach 1: Random walks / IP addresses
  - In theory: might give us a true estimate of the size of the web (as opposed to just relative sizes of indexes)
- Approach 2: Generate a random URL contained in a given engine
  - Suffices for accurate estimation of relative size

Statistical methods

1. Random queries
2. Random searches
3. Random IP addresses
4. Random walks

Random URLs from random queries

1. Generate random query: how?

   Lexicon: 400,000+ words from a web crawl

   Conjunctive Queries: w_1 and w_2

   e.g., vocalists AND rsi

   Not an English dictionary

2. Get 100 result URLs from engine A
3. Choose a random URL as the candidate to check for presence in engine B

- This distribution induces a probability weight W(p) for each page.

Query Based Checking

- Either *search for the URL* if the engine B support this or
- *Generate a Strong Query* to check whether an engine *B* has a document *D*:
  - Download *D*. Get list of words.
  - Use 8 low frequency words as AND query to *B*
  - Check if *D* is present in result set.

Advantages & disadvantages

- Statistically sound under the induced weight.
- Biases induced by random query
  - Query Bias: Favors content-rich pages in the language(s) of the lexicon
  - Ranking Bias: *Solution:* Use conjunctive queries & fetch all (picking from top 100)
  - Checking Bias: Duplicates, impoverished pages omitted
  - Document or query restriction bias: engine might not deal properly with 8 words conjunctive query
  - Malicious Bias: Sabotage by engine
  - Operational Problems: Time-outs, failures, engine inconsistencies, index modification.

Random searches

- Choose random searches extracted from a local query log [Lawrence & Giles 97] or build "random searches" [Notess]

- Use only queries with small result sets.
- For each random query: compute ratio size(r_1)/size(r_2) of the two result sets
- Average over random searches

Advantages & disadvantages

- Advantage
  - Might be a better reflection of the human perception of coverage
- Issues
  - Samples are correlated with source of log (unfair advantage for originating search engine)
  - Duplicates
  - Technical statistical problems (must have non-zero results, ratio average not statistically sound)

Random searches

- 575 & 1050 queries from the NEC RI employee logs
- 6 Engines in 1998, 11 in 1999
- Implementation:
  - Restricted to queries with < 600 results in total
  - Counted URLs from each engine after verifying query match
  - Computed size ratio & overlap for individual queries
  - Estimated index size ratio & overlap by averaging over all queries

Queries from Lawrence and Giles study

- *adaptive access control*
- *neighborhood preservation topographic*
- *hamiltonian structures*
- *right linear grammar*
- *pulse width modulation neural*
- *unbalanced prior probabilities*
- *ranked assignment method*
- *internet explorer favourites importing*
- *karvel thornber*
- *zili liu*
- *softmax activation function*
- *bose multidimensional system theory*
- *gamma mlp*
- *dvi2pdf*
- *john oliensis*
- *rieke spikes exploring neural*
- *video watermarking*
- *counterpropagation network*
- *fat shattering dimension*
- *abelson amorphous computing*

Random IP addresses

- Generate random IP addresses
- Find a web server at the given address
  - If there's one
- Collect all pages from server
  - From this, choose a page at random

Random IP addresses

- HTTP requests to random IP addresses
  - Ignored: empty or authorization required or excluded
  - [Lawr99] Estimated 2.8 million IP addresses running crawlable web servers (16 million total) from observing 2500 servers.
  - OCLC using IP sampling found 8.7 M hosts in 2001
    - Netcraft [Netc02] accessed 37.2 million hosts in July 2002
- [Lawr99] exhaustively crawled 2500 servers and extrapolated
  - Estimated size of the web to be 800 million pages
  - Estimated use of metadata descriptors:
    - Meta tags (keywords, description) in 34% of home pages, Dublin core metadata in 0.3%

Advantages & disadvantages

- Advantages
  - Clean statistics
  - Independent of crawling strategies
- Disadvantages
  - Doesn't deal with duplication
  - Many hosts might share one IP (oversampling), or not accept requests
  - No guarantee all pages are linked to root page.
    - E.g.: employee pages
  - Power law for # pages/hosts generates bias towards sites with few pages. (under-sampling)
    - But bias can be accurately quantified IF underlying distribution understood
  - Potentially influenced by spamming (multiple IP's for same server to avoid IP block)

Τυχαίοι Περίπατοι (Random walks)

Το διαδίκτυο ως ένας κατευθυνόμενος

- Ένας τυχαίος περίπατος σε αυτό το γράφο
  - Includes various "jump" rules back to visited sites
    - Does not get stuck in spider traps!
    - Can follow all links!
  - Συγκλίνει σε μια κατανομή σταθερής κατάστασης (stationary distribution)
    - Must assume graph is finite and independent of the walk.
    - Conditions are not satisfied (cookie crumbs, flooding)
    - Time to convergence not really known
  - Sample from stationary distribution of walk
  - Use the "strong query" method to check coverage by SE

Advantages & disadvantages

- Advantages
  - "Statistically clean" method, at least in theory!
  - Could work even for infinite web (assuming convergence) under certain metrics.
- Disadvantages
  - List of seeds is a problem.
  - Practical approximation might not be valid.
  - Non-uniform distribution
    - Subject to link spamming

Size of the web

Check out
http://www.worldwidewebsize.com/

The Indexed Web contains **at least 3.57 billion pages** (Tuesday, 20 May, 2014).

Conclusions

- No sampling solution is perfect.
- Lots of new ideas but the problem is getting harder

ΤΕΛΟΣ 10ου Μαθήματος

Ερωτήσεις?

Χρησιμοποιήθηκε κάποιο υλικό από:
- ✓*Pandu Nayak and Prabhakar Raghavan, CS276:Information Retrieval and Web Search (Stanford)*
- ✓*Hinrich Schütze and Christina Lioma, Stuttgart IIR class*