Introduction to

Information Retrieval

ΠΛΕ70: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Διάλεξη 9: Βασικές Αρχές Αναζήτησης στον Παγκόσμιο Ιστό.

1

Τι θα δούμε σήμερα;

- Ιστορικά στοιχεία και γενικές πληροφορίες
- Πόσο μεγάλος είναι ο Ιστός;
- Διαφημίσεις, spam
- Διπλότυπες σελίδες

2

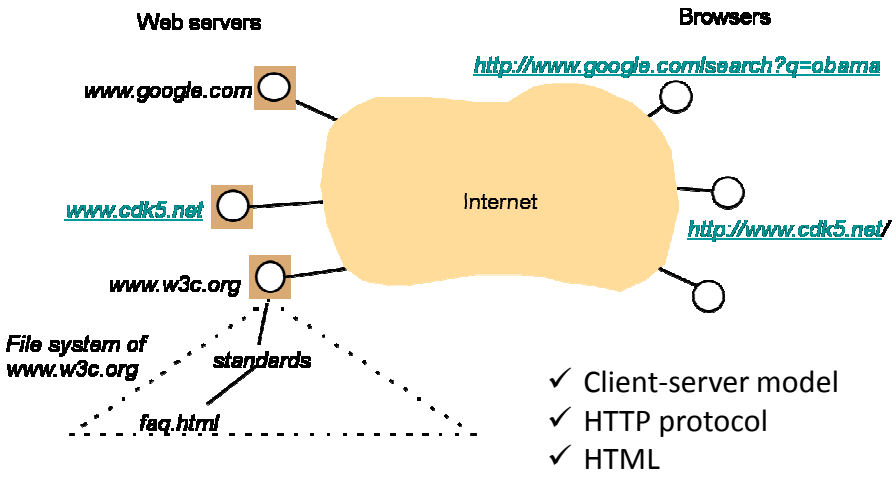
Web (WWW)

World Wide Web (World-Wide Web, **WWW**, **W3**, ή **the Web**) **είναι μια συλλογή από έγγραφα κειμένου και άλλες πηγές** (**web σελίδες - ιστοσελίδες**), που είναι συνδεδεμένα hyperlinks και URLs,

- hosted **web servers**
- viewed or navigated via hyperlinks with **web browsers**.

- 63 billion pages
- 1 trillion unique web addresses

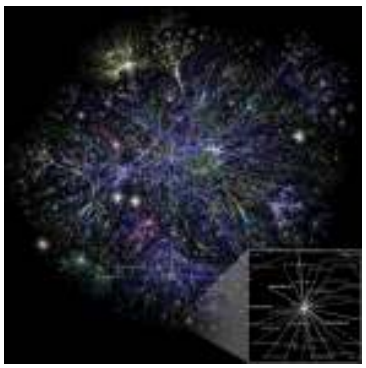
Web (WWW): Function



- ✓ Client-server model
- ✓ HTTP protocol
- ✓ HTML

Internet

Το διαδίκτυο (Internet) είναι ολικό σύστημα δια-συνδεδεμένων δικτύων υπολογιστών που χρησιμοποιούν ένα standard Internet protocol suite (TCP/IP)



Το Web είναι μια εφαρμογή που τρέχει πάνω στο Internet

Web (WWW): Function

Viewing a web page

- either by *typing the URL* of the page into a web browser or
- by *following a hyperlink* to that page or resource.

The web browser then initiates a series of communication messages, to **fetch** and **display** it.

Web (WWW): Function

URL http://en.wikipedia.org/wiki/World_Wide_Web .

1. Browser resolves the server-name portion of the URL into an IP (Internet Protocol) address using the globally distributed database known as the *Domain Name System (DNS)*
[returns an IP address such as *208.80.152.2*.]

URL – (DNS) -> IP address

Web (WWW): Function

URL http://en.wikipedia.org/wiki/World_Wide_Web .

Domain/file-under-the-root-directory of the server

2. Browser then requests the resource by sending an HTTP request across the Internet to the computer at that particular address.
It makes the request to a particular application port in the underlying IP normally port 80. The content of the HTTP request can be as simple as the two lines of text

GET */wiki/World_Wide_Web* HTTP/1.1 Host: *en.wikipedia.org*

Web (WWW): Function

1. The computer receiving the HTTP request delivers it to Web server software listening for requests on port 80.

2. If the web server can fulfill the request it sends an HTTP response back to the browser indicating success, which can be as simple as

```
HTTP/1.0 200 OK Content-Type: text/html; charset=UTF-8
```

followed by the content of the requested page.

Web (WWW): Function

The Hypertext Markup Language for a basic web page

```
<html>
<head>
<title> World Wide Web — Wikipedia, the free encyclopedia </title>
</head>
<body> <p>
The World Wide Web, abbreviated as WWW and commonly known ...</p> </body>
</html>
```

The web browser parses the HTML, interpreting the markup (<title>, <p> for paragraph, and such) to draw that text on the screen.

✓ Ignores what it cannot understand

Web (WWW): Function

- Many web pages consist of *more elaborate HTML* which references the URLs of other resources such as images, other embedded media, scripts that affect page behavior, and *Cascading Style Sheets* that affect page layout.
- (Asynchronous) A browser that handles complex HTML will make *additional HTTP requests* to the web server for these other Internet media types.
- As it receives their content from the web server, the browser progressively renders the page onto the screen as specified by its HTML and these additional resources.

Web (WWW): Linking

Most web pages contain hyperlinks to other related pages and perhaps to downloadable files, source documents, definitions and other web resources

In the underlying HTML, a hyperlink looks like

```
<a href="http://www.w3.org/History/19921103hypertext/hypertext/WWW/">Early archive of the first Web site</a>
```

The hyperlink structure of the WWW described by the web graph

Web (WWW): Ιστορία

Στο τεύχος του **Ιουνίου 1970** του περιοδικού *Popular Science*

Arthur C. Clarke
satellites would one day "bring the accumulated knowledge of the world to your fingertips" using a console that would combine the functionality of the Xerox, telephone, television and a small computer, allowing data transfer and video conferencing around the globe.

Web (WWW): History

1980, **Tim Berners-Lee** a proposal that referenced ENQUIRE, a database and software project he had built in 1980

November 1990, with ***Robert Cailliau***, a more formal proposal to build a "Hypertext project" called "WorldWideWeb" (one word, also "W3") as *a "web" of "hypertext documents" to be viewed by "browsers" using a client–server architecture.*

Estimated that *a read-only web* would be developed *within 3 months* and that it would take *6 months* to achieve "the creation of *new links and new material by readers*, [so that]

"authorship becomes universal" as well as "the *automatic notification* of a reader when new material of interest to him/her has become available."

Web (WWW): History

By Christmas 1990, *all tools* for a working Web:

- the first web browser (which was a web editor as well);
- the first web server and
- the first web pages, which described the project itself.

August 6, 1991, post on alt.hypertext newsgroup -> the debut of the Web as a publicly available service on the Internet.

Web (WWW): History

Ο πρώτος web server (και πρώτος web browser): A NeXT Computer -

Η πρώτη φωτογραφία στο web το 1992 (CERN house band Les Horribles Cernettes)

The Web's historic logo designed by Robert Cailliau

Web (WWW): History, why in CERN?

Web as a "Side Effect" of the 40 years of Particle Physics Experiments.

After the World War 2. the nuclear centers of almost all developed countries became the places with the highest concentration of talented scientists.

For about four decades many of them were invited to the international CERN's Laboratories.

Web (WWW): History

Berners-Lee's breakthrough: marry hypertext to the Internet

3 essential technologies:

1. a system of globally unique identifiers for resources on the Web and elsewhere, the Universal Document Identifier (UDI), later known as Uniform Resource Locator (**URL**) and Uniform Resource Identifier (URI);
2. the publishing language HyperText Markup Language (**HTML**);
3. the Hypertext Transfer Protocol (**HTTP**)

Web (WWW): History

Differences from other hypertext systems

❖ required only *unidirectional links* rather than bidirectional ones.

(+) possible for someone to link to another resource without action by the owner of that resource
(+) reduced the difficulty of implementing web servers and browsers (in comparison to earlier systems)
(-) presented the chronic problem of *link rot (or dead links)*.

❖ was *non-proprietary* (unlike, e.g., HyperCard)

making it possible to develop servers and clients independently and to add extensions without licensing restrictions.

On April 30, 1993, CERN announced that the World Wide Web would be free to anyone, with no fees due. Coming two months after the announcement that the server implementation of the Gopher protocol was no longer free to use, this produced a rapid shift away from Gopher and towards the Web.

Web (WWW): History

Early popular web browser was ViolaWWW for Unix and the X Windowing System.

In **1993**, **Mosaic** web browser, a *graphical browser* developed by a team at the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign (NCSA-UIUC), led by Marc Andreessen.

Prior to the release of Mosaic, graphics were not commonly mixed with text in web pages

Web (WWW): History

The **World Wide Web Consortium** (W3C) was founded by Tim Berners-Lee after he left (CERN) in October 1994.

World Wide Web and *Internet*

Web2.0

A **Web 2.0 site** allows users to *interact* and *collaborate* **(prosumers)** of user-generated content in a virtual community,

Examples of Web 2.0 include social networking sites, blogs, wikis, video sharing sites, hosted services, web applications, mashups and folksonomies.

Web.2: History

The term "Web 2.0" was first used in **January 1999** by **Darcy DiNucci**, a consultant on electronic information design (information architecture). In her article, "Fragmented Future", DiNucci writes:

The Web we know now, which loads into a browser window in essentially static screenfuls, is only an embryo of the Web to come. The first glimmerings of Web 2.0 are beginning to appear, and we are just starting to see how that embryo might develop.
The Web will be understood not as screenfuls of text and graphics but as a transport mechanism, the ether through which interactivity happens. It will [...] appear on your computer screen, [...] on your TV set [...] your car dashboard [...] your cell phone [...] hand-held game machines [...] maybe even your microwave oven.

Web.2: History

In 2003, rise in popularity when **O'Reilly Media** and MediaLive hosted the first Web 2.0 conference.

In their opening remarks, John Battelle and Tim O'Reilly outlined their definition of the "Web as Platform", where software applications are built upon the Web as opposed to upon the desktop.

Web.2: History

In the **2006** ,TIME magazine Person of The Year (You).

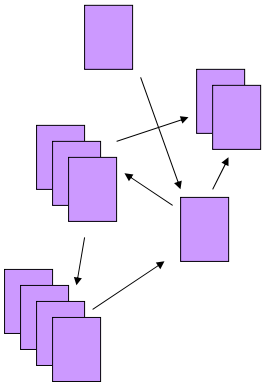
TIME selected the masses of users who were participating in content creation on social networks, blogs, wikis, and media sharing sites.

In the cover story, Lev Grossman:

It's a story about community and collaboration on a scale never seen before. It's about the cosmic compendium of knowledge Wikipedia and the million-channel people's network YouTube and the online metropolis MySpace. It's about the many wresting power from the few and helping one another for nothing and how that will not only change the world but also change the way the world changes.

In **2009,** Global Language Monitor declare Web2.0 to be the one-millionth English word

The Web document collection



The Web

- No design/co-ordination
- Distributed content creation, linking, democratization of publishing
- Content includes truth, lies, obsolete information, contradictions …
- Unstructured (text, html, …), semi-structured (XML, annotated photos), structured (Databases)…
- Scale much larger than previous text collections … but corporate records are catching up
- Growth – slowed down from initial "volume doubling every few months" but still expanding
- Content can be *dynamically generated*

Search Engines

- *Full text search* (Altavista, Excite, Infoseek)
- *Taxonomies* (Yahoo!) – browse through a hierarchical tree with category labels

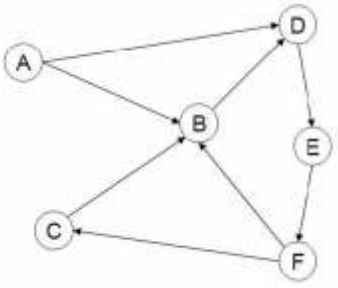
About.com Open Directory Project

Dynamic vs static web pages

✓ Hidden web – Deep web
✓ Personal web site vs airport flight status

URL: not a file but a program on the server
Input part of the GET, e.g., http//www.google.com/search?q=obama

The Web graph



Anchor text <a>
In-links/Out-links
In-degree (8-15)
Out-degree

The Web Graph

- the distribution of in-degrees not Poisson distribution (if every web page were to pick the destinations of its links uniformly at random).

- Power law,

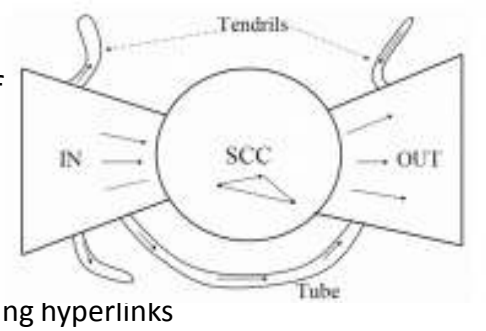
the total number of web pages with in-degree ***i*** is proportional to $1/i^{\alpha}$

α typically 2.1

The Web graph

Bowtie shape

Three major categories of web pages
IN, **OUT**, **SCC**



A web surfer can pass by following hyperlinks

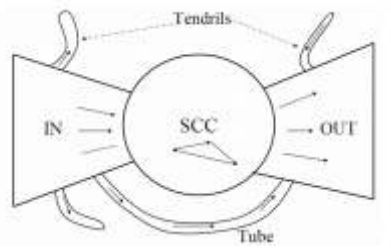
- from any page in IN to any page in SCC,
- from any page in SCC to any page in OUT.
- from any page in SCC to any other page in SCC.
- not possible to pass from a page in SCC to any page in IN, a page in OUT to a page in SCC (or, consequently, IN).

31

The Web graph

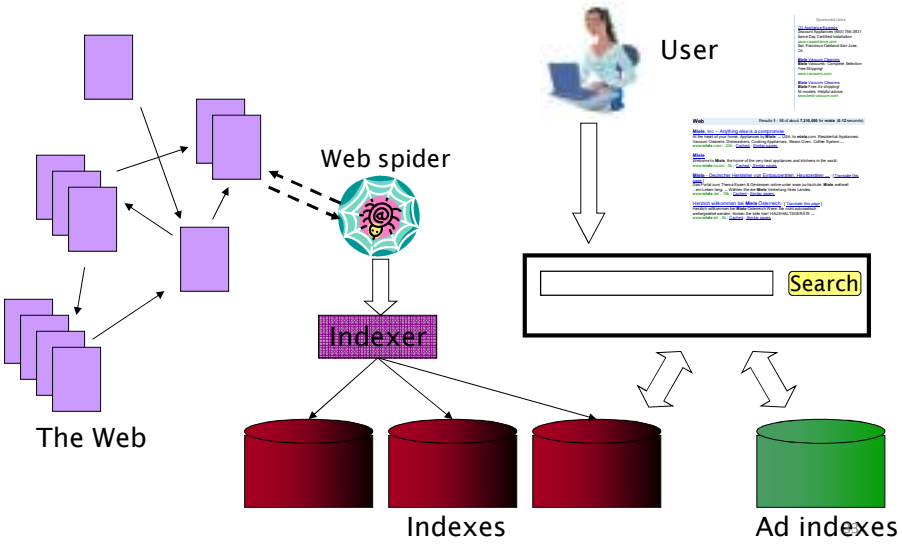
IN, OUT same size, SCC larger
Remaining pages:

- Tubes: small sets of pages outside SCC that lead directly from IN to OUT,
- Tendrils: either lead nowhere from IN, or from nowhere to OUT.



32

Web search basics



ΟΙ ΧΡΗΣΤΕΣ

Ανάγκες Χρηστών

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

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 ίση με 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

Seattle weather

Mars surface images

Canon S410

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

Car rental Brasil

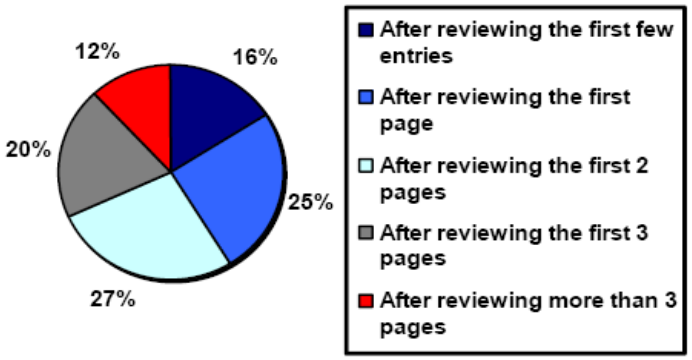
Ανάγκες Χρηστών

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)

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

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 would 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

Graphical graph banners on popular web sites (branding)

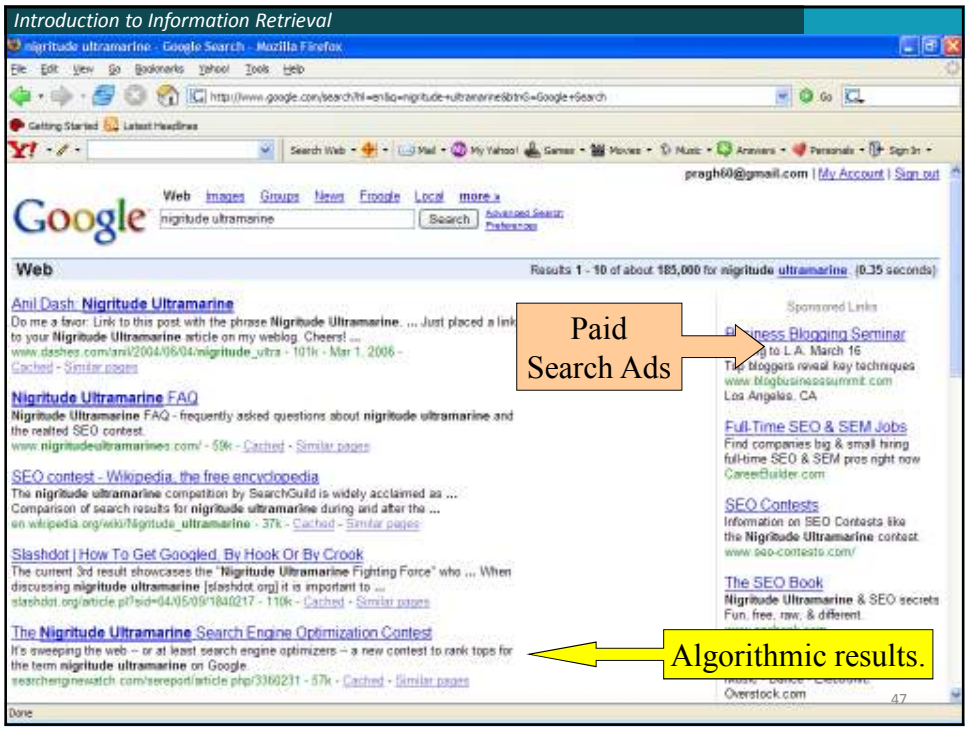
- ***cost per mil (CPM) model***: the cost of having its banner advertisement displayed 1000 times (also known as impressions)
- ***cost per click (CPC) model***: number of clicks on the advertisement (leads to a web page set up to make a purchase)

✓ brand promotion vs and transaction-oriented advertising

Ads

Provide

- ***pure search results*** (generally known as algorithmic 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.



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.

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

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.

Exercise

- Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?
- The advertiser pays for all this. How can the advertiser be cheated?
- Any way this could be bad for the user?
- Any way this could be bad for the search engine?

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.

55

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.
- See http://google.com/tm complaint.html
- 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

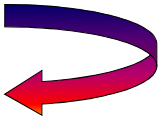
Search engine optimization (Spam)

- Motives
  - Commercial, political, religious, lobbies
  - Promotion funded by advertising budget
- Operators
  - Contractors (Search Engine Optimizers) for lobbies, companies
  - Web masters
  - Hosting services
- Forums
  - E.g., Web master world (www.webmasterworld.com)
    - Search engine specific tricks
    - Discussions about academic papers ☺

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»

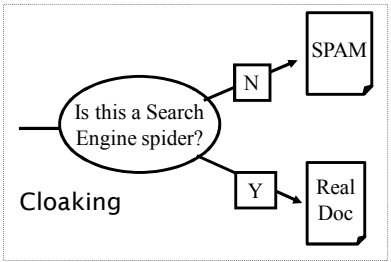
a web page loaded with keywords in the meta tags
or in content of a web page (outdated)

- 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.
- **Link spamming**
  - Mutual admiration societies, hidden links, awards – more on these later
  - *Domain flooding:* numerous domains that point or re-direct to a target page
- **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;

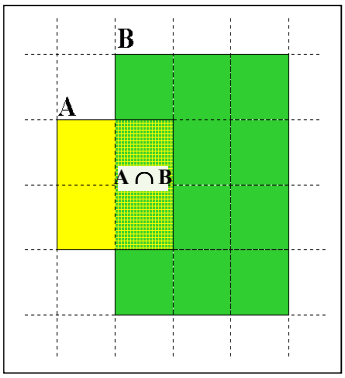
Το σχετικό μέγεθος των μηχανών αναζήτησης

- 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

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

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

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

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

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

Δειγματοληψία (Sampling) URLs

Ιδανική στρατηγική: Παρήγαγε ένα τυχαίο URL και έλεγξε αν περιλαμβάνετε σε κάθε ευρετήριο.

- Problem: Random URLs are hard to find! Enough to generate a random URL contained in a given Engine.
- Approach 1: Generate a random URL contained in a given engine
  - Suffices for the estimation of relative size
- Approach 2: 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)

Statistical methods

- Approach 1
  - Random queries
  - Random searches
- Approach 2
  - Random IP addresses
  - Random walks

Random URLs from random queries

- Generate random query: how?
  - **Lexicon:** 400,000+ words from a web crawl
  - **Conjunctive Queries:** w_1 and w_2
    e.g., vocalists AND rsi
- Get 100 result URLs from engine A
- 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.

Not an English dictionary

Query Based Checking

- *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
  - 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.
  - Count normalized URLs in result sets.
  - Use ratio statistics

Advantages & disadvantages

- Advantage
  - Might be a better reflection of the human perception of coverage
- Issues
  - Samples are correlated with source of log
  - 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, 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.
    - 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/

Conclusions

- No sampling solution is perfect.
- Lots of new ideas ...
- ....but the problem is getting harder
- Quantitative studies are fascinating and a good research problem

DUPLICATE DETECTION

Duplicate documents

- The web is full of duplicated content
- Strict duplicate detection = exact match
  - Not as common
- But many, many cases of near duplicates
  - E.g., last-modified date the only difference between two copies of a page

Duplicate/Near-Duplicate Detection

- *Duplication*: Exact match can be detected with fingerprints
- *Near-Duplication*: Approximate match
  - Overview
    - Compute syntactic similarity with an edit-distance measure
    - Use similarity threshold to detect near-duplicates
      - E.g., Similarity > 80% => Documents are "near duplicates"
      - Not transitive though sometimes used transitively

Computing Similarity

- Features:
  - Segments of a document (natural or artificial breakpoints)
  - Shingles (Word N-Grams)
  - ***a rose is a rose is a rose*** →

    a_rose_is_a

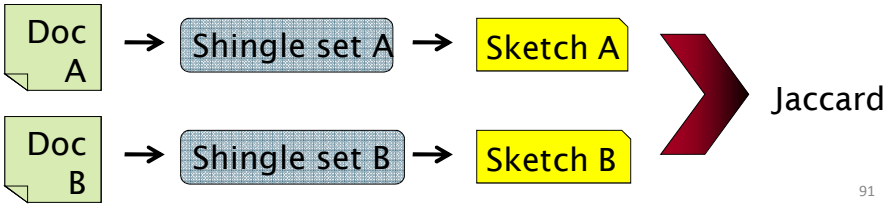
    rose_is_a_rose

    is_a_rose_is

    a_rose_is_a
- Similarity Measure between two docs (= sets of shingles)
  - Jaccard coefficient: Size_of_Intersection / Size_of_Union

Shingles + Set Intersection

- Computing exact set intersection of shingles between all pairs of documents is expensive/intractable
  - Approximate using a cleverly chosen subset of shingles from each (a *sketch*)
- Estimate (size_of_intersection / size_of_union) based on a short sketch



ΤΕΛΟΣ 9ου 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*