ΜΥΕ003: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Κεφάλαια 19, 20: Web. Μηχανές αναζήτησης.

Τι θα δούμε σήμερα

- Web και μηχανές αναζήτησης
  - λίγη ιστορία
  - ο γράφος του web
  - σημασία της άγκυρας (anchor text)
- Χρήστες
- Εξατομίκευση
- Διαφημίσεις

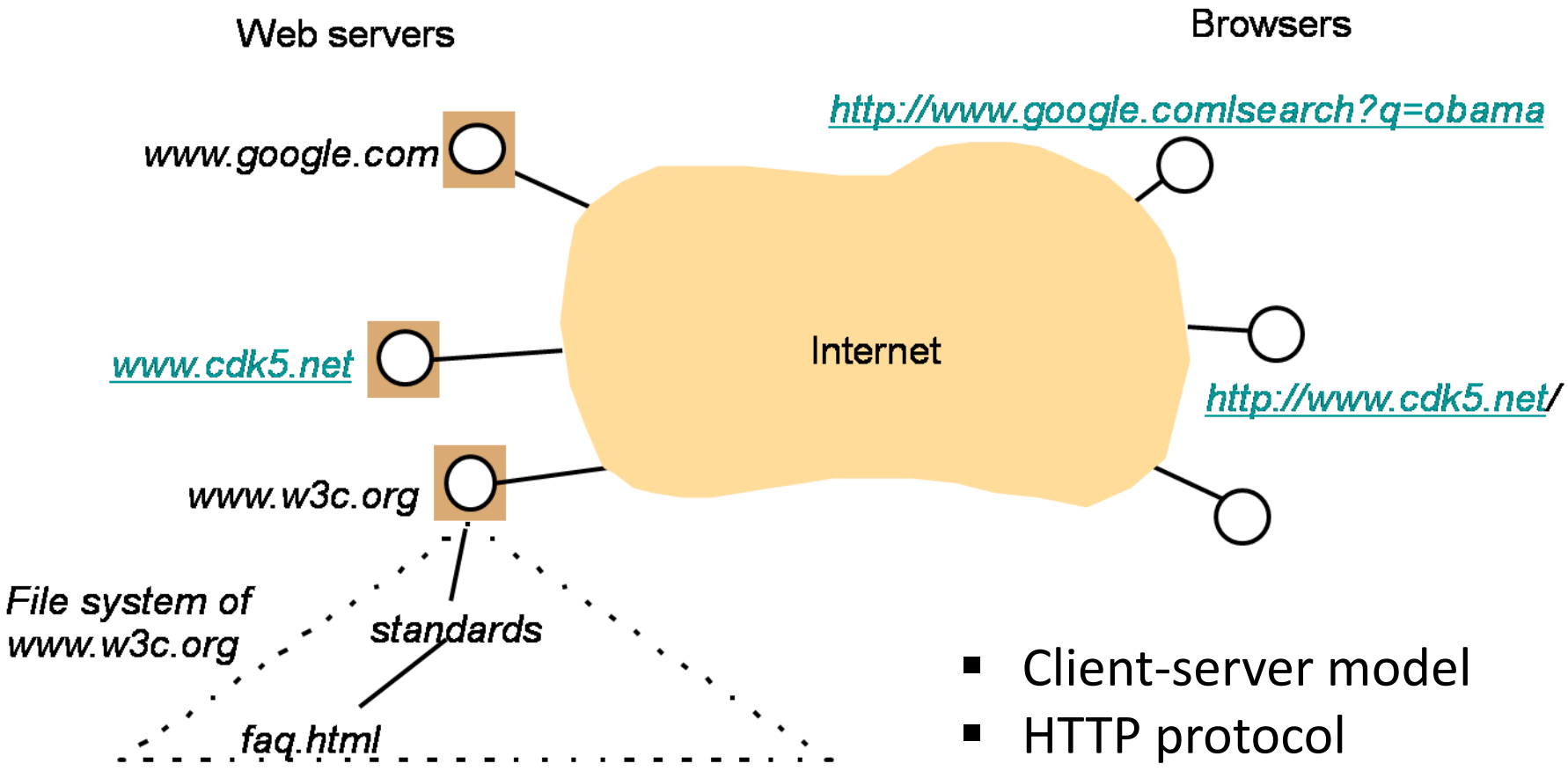
Web: τι είναι

Web (World Wide Web, WWW, W3)

- μια συλλογή από *web σελίδες* (ιστοσελίδες) που είναι έγγραφα κειμένου και άλλες πηγές συνδεδεμένα με *hyperlinks* και *URLs*
- μια εφαρμογή που τρέχει πάνω από το Internet

- 63 δισεκατομμύρια ιστοσελίδες
- 1 τρισεκατομμύριο διαφορετικές web διευθύνσεις

Web: η δομή του



- Client-server model
- HTTP protocol
- HTML
- URL/URI

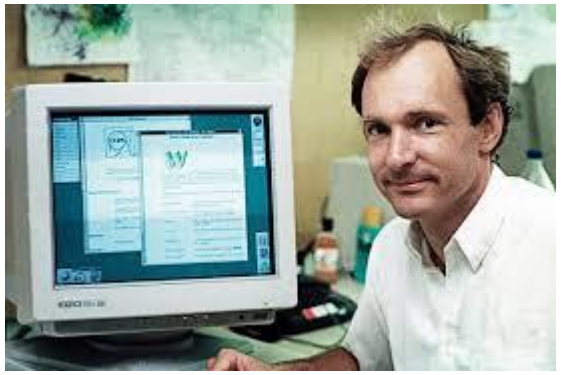
Web: Ιστορία

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: Ιστορία

1980, Tim Berners-Lee (ENQUIRE)



Νοέμβριο 1990, με τον *Robert Cailliau*, πρόταση για ένα "Hypertext project με το όνομα "WorldWideWeb" ("W3"): *"web" of "hypertext documents" to be viewed by "browsers" using a client–server architecture.*

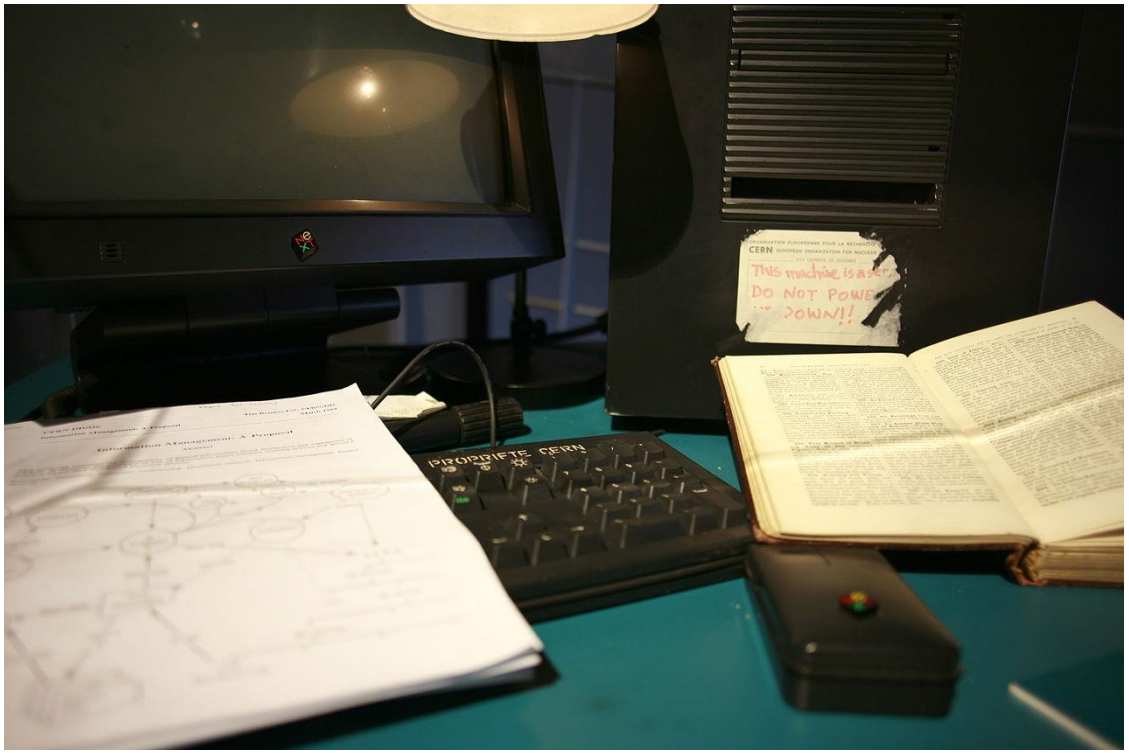
Χριστούγεννα 1990, το πρώτο λειτουργικό σύστημα:

- ο πρώτος web browser (που ήταν και web editor),
- ο πρώτος web server, και
- οι πρώτες ιστοσελίδες, που περιέγραφαν το ίδιο το project.

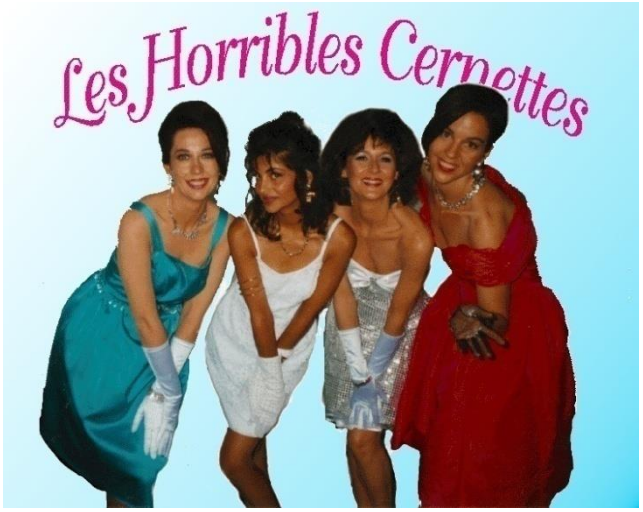
Αύγουστο 1991, post στο alt.hypertext newsgroup – νέο service στο ‘Ιντερνετ

Web: Ιστορία

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



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



logo by Robert Cailliau

Mosaic (1993) πρώτος graphical browser

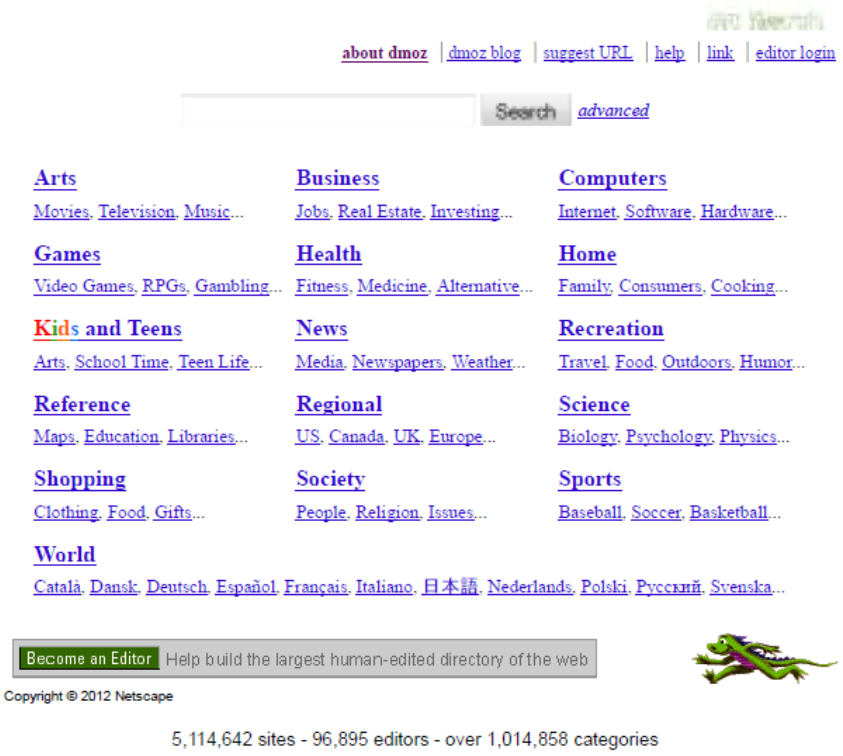
Web: Ιστορία

2016 ACM Turing Award
Time's magazine list of the 100 most influential people of the 20th century

Εύρεση Πληροφορίας

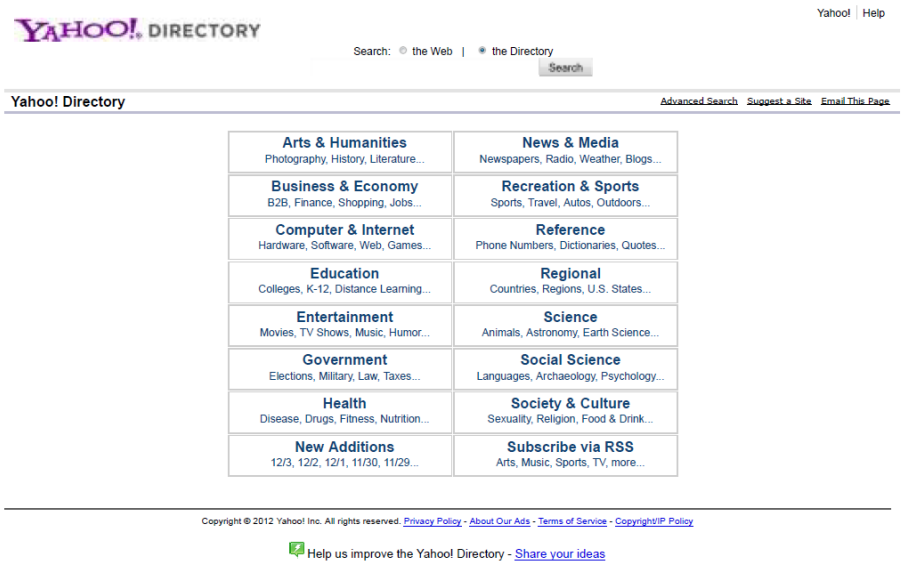
Πρώτη προσπάθεια: Ευρετήρια (κατάλογοι) που τα διατηρούν άνθρωποι (μη αυτοματοποιημένα)

DMOZ - Open Directory Project



As of Mar 17, 2017, dmoz.org is no longer available.

Εύρεση Πληροφορίας



officially closed on December 31, 2014

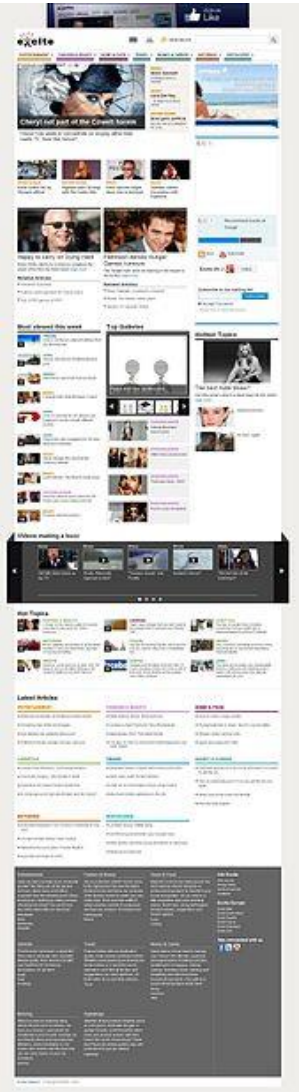
Εύρεση Πληροφορίας

Δεύτερη προσπάθεια: Μηχανές αναζήτησης πλήρους κειμένου (full text)

Πόσο συχνά ο όρος εμφανίζεται στο κείμενο

Αρχικά οι χρήστες τις σελίδες τους για να ευρετηριοποιηθούν

Altavista



Excite

infoseek
Once you know, you know.
BigYellow

Infoseek

Type a specific question, phrase or Name.
the Web
seek
Tips
To explore the Web's largest directory, click a topic below.

Εύρεση Πληροφορίας

- Η εποχή του Google (~1998): χρήση του web ως γράφου
  - Πέρασμα από τη *συνάφεια* στο *κύρος* (authoritativeness)
  - Δεν έχει μόνο σημασία μια σελίδα να είναι συναφής πρέπει να είναι και *σημαντική* στο web

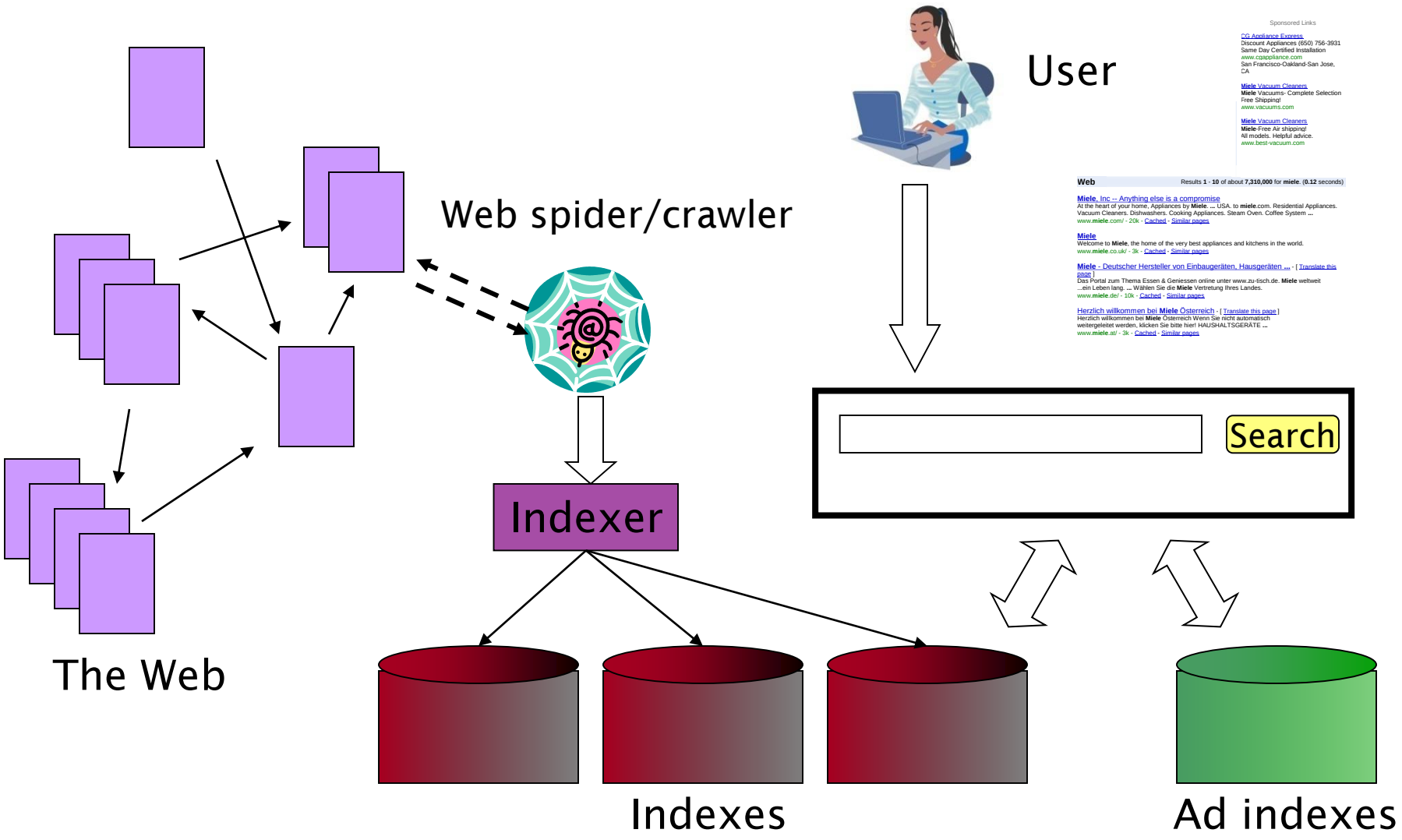
Larry Page, Sergey Brin

Εύρεση Πληροφορίας

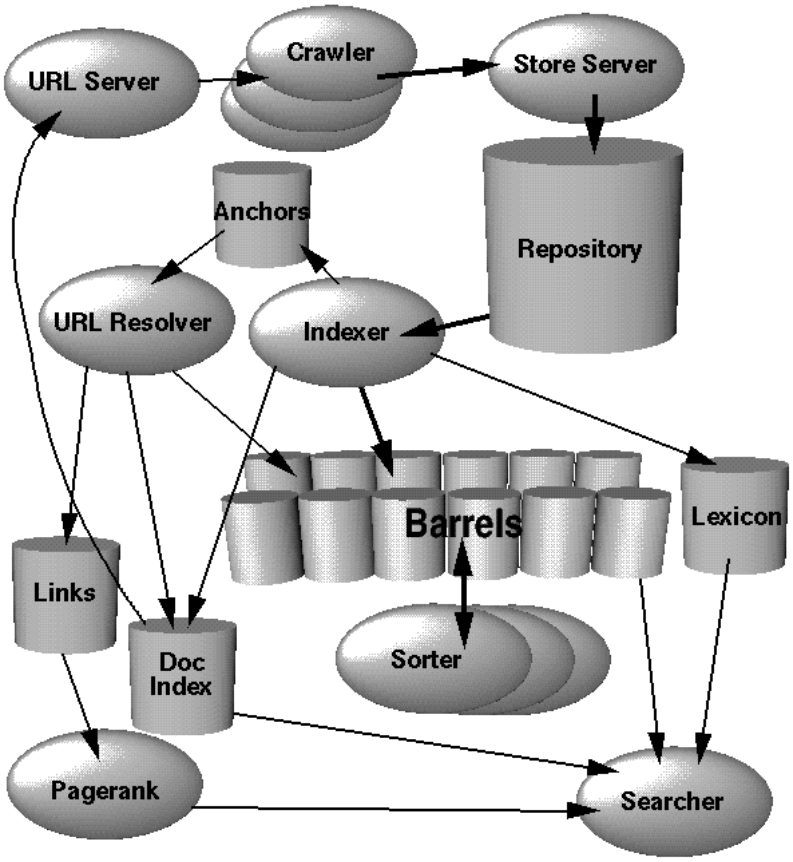
από τη λέξη googol, ο αριθμός 1 ακολουθούμενος από 100 μηδενικά

αρχικά στο website του Stanford University

Αναζήτηση στο Web: βασικά σημεία



Search Engine Anatomy*



BigFiles: virtual files spanning multiple file systems
Repository: full HTML of every web page
DocIndex: info about files

Hit list: list of occurrences of a particular word in a particular document including position, font, and capitalization (fancy (title, anchor, etc), plain)

Forward index sorter -> Inverted index
Barrel: ranges of wordids (forward, backward barrels)

** The Anatomy of a Large-Scale Hypertextual Web Search Engine, Sergey Brin and Lawrence Page*

Δυναμικές και στατικές σελίδες

Στατικές: σελίδες που το περιεχόμενο τους δεν αλλάζει από την μία αίτηση στην άλλη

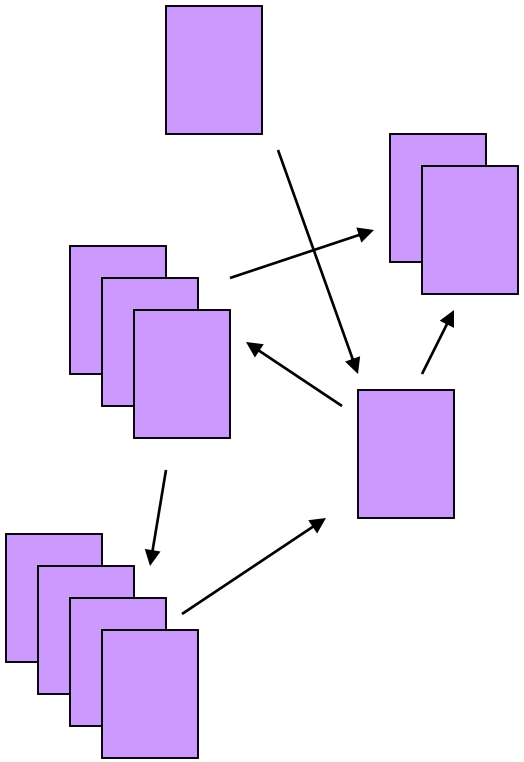
Δυναμικές σελίδες: Hidden web – Deep web

- ✓ Παράδειγμα: προσωπική ιστοσελίδα vs σελίδα με την κατάσταση των πτήσεων σε ένα αεροδρόμιο

URL: συνήθως όχι κάποιο αρχείο αλλά κάποιο πρόγραμμα στον server

Input part of the GET, e.g., http//www.google.com/search?q=obama

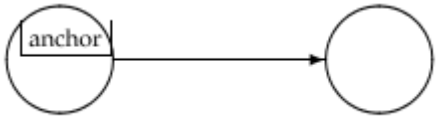
Η συλλογή εγγράφων του Web



The Web

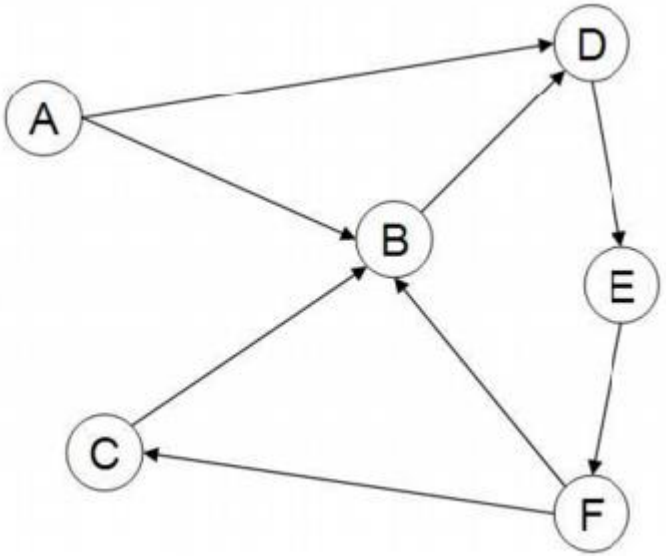
- No design/co-ordination
- *Distributed* content creation, linking, democratization of publishing
- Content includes *truth, lies*, obsolete information, contradictions …
- *Unstructured* (text, html, …), semi-structured (XML, annotated photos), structured (Databases)…
- *Scale* much larger than previous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

Ο γράφος του Web



Anchor text <a>

- Κατευθυνόμενος, (In-links/Out-links)
- In-degree (8-15)
- Out-degree



Ο γράφος του Web

- Η κατανομή των εισερχόμενων ακμών *δεν* ακολουθεί την κατανομή Poisson (την κατανομή που θα είχαμε αν κάθε σελίδα επέλεγε ποιες σελίδες θα κάνει link τυχαία (uniformly at random).

- Κατανομή Power law,

Το πλήθος των σελίδων με in-degree i είναι ανάλογο του $1/i^{\alpha}$

α είναι συνήθων 2,1

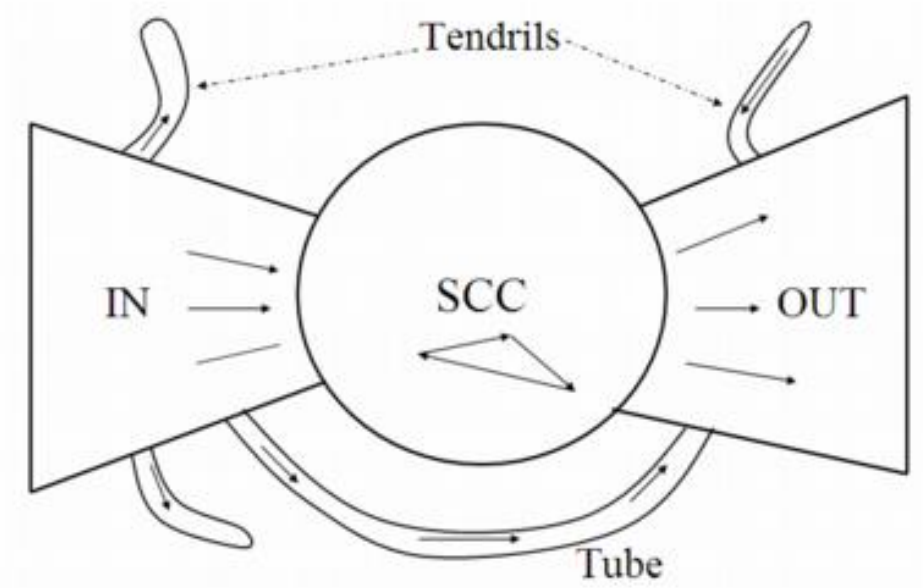
Που αλλού είδαμε παρόμοια κατανομή;

Ο γράφος του Web

Bow-tie shape

Τρεις κατηγορίες: **IN**, **OUT**, **SCC**

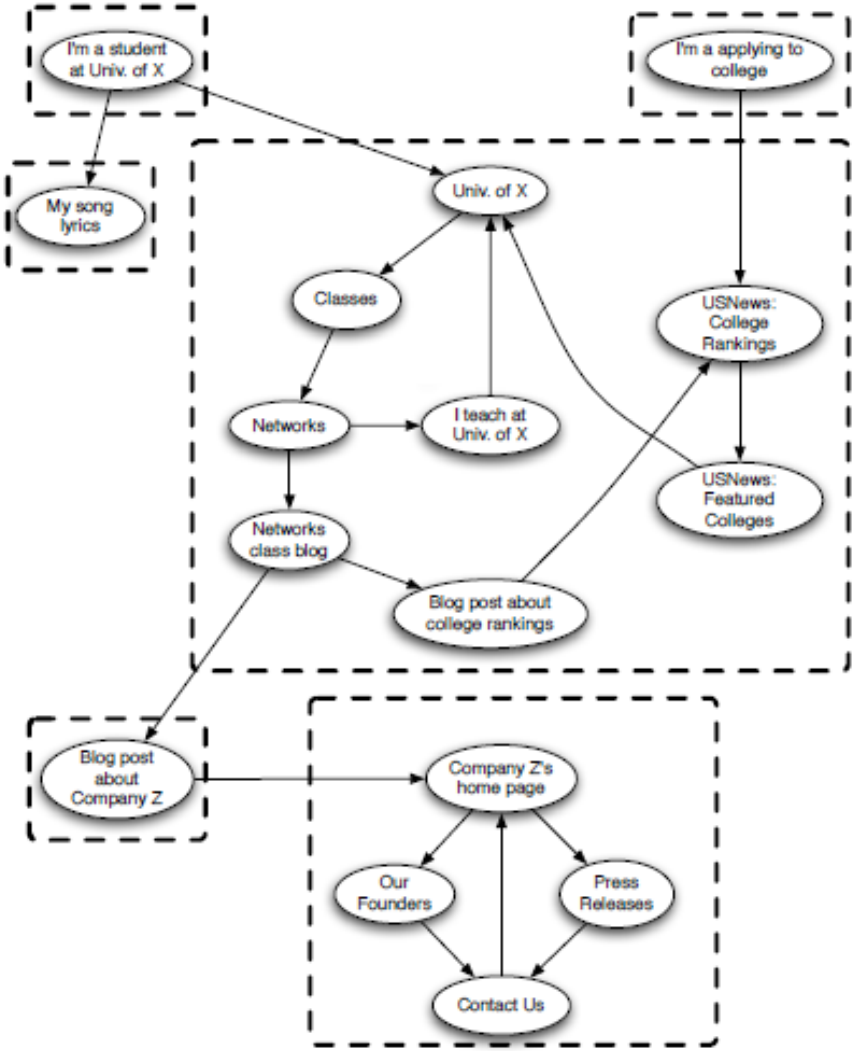
Περιέχει μια μεγάλη ισχυρά συνδεδεμένη συνιστώσα (Strongly Connected Component (SCC))



IN: Σελίδες που οδηγούν στο SCC αλλά όχι το ανάποδο

OUT: Σελίδες στις οποίες μπορούμε να φτάσουμε από το SCC αλλά δεν οδηγούν σε αυτό

Ο γράφος του Web



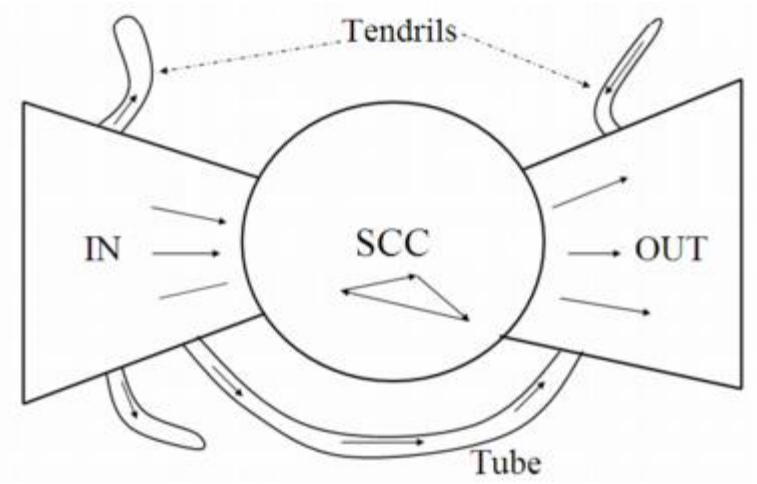
From the book Networks, Crowds, and Markets: Reasoning ab
Highly Connected World. By David Easley and Jon Kleinberg. C
University Press, 2010. Complete preprint on-line at
http://www.cs.cornell.edu/home/kleinber/networks-book/

Ο γράφος του Web

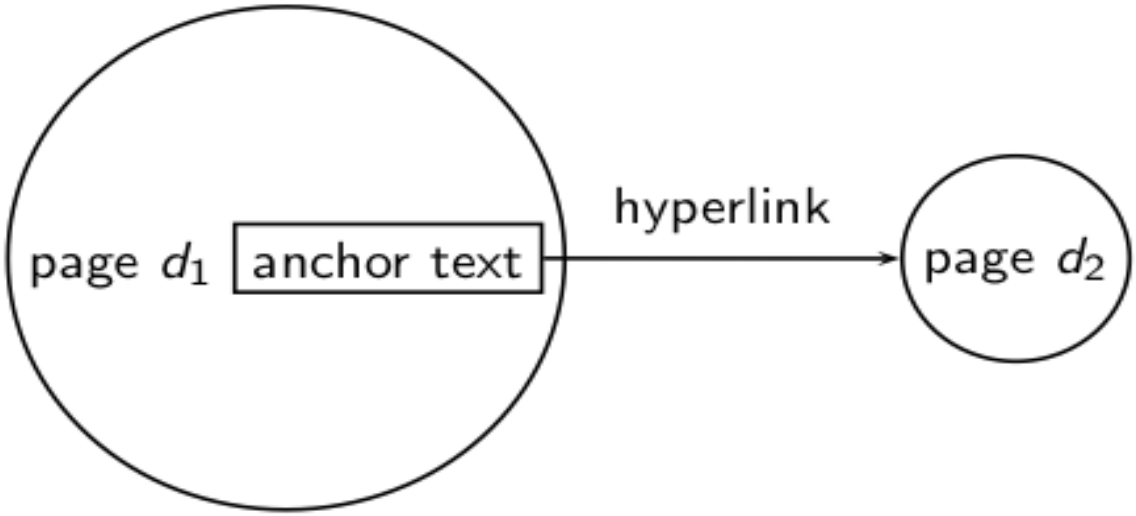
IN, OUT παρόμοιο μέγεθος, SCC μεγαλύτερο
Υπόλοιπες σελίδες:

- Tubes: μικρά σύνολα σελίδων έξω από το SCC που δείχνουν απευθείας από το IN στο OUT,
- Tendrils: είτε δεν οδηγούν πουθενά από το IN είτε δείχνουν από πουθενά στο OUT.

Μικρές μη συνδεδεμένες συνιστώσες



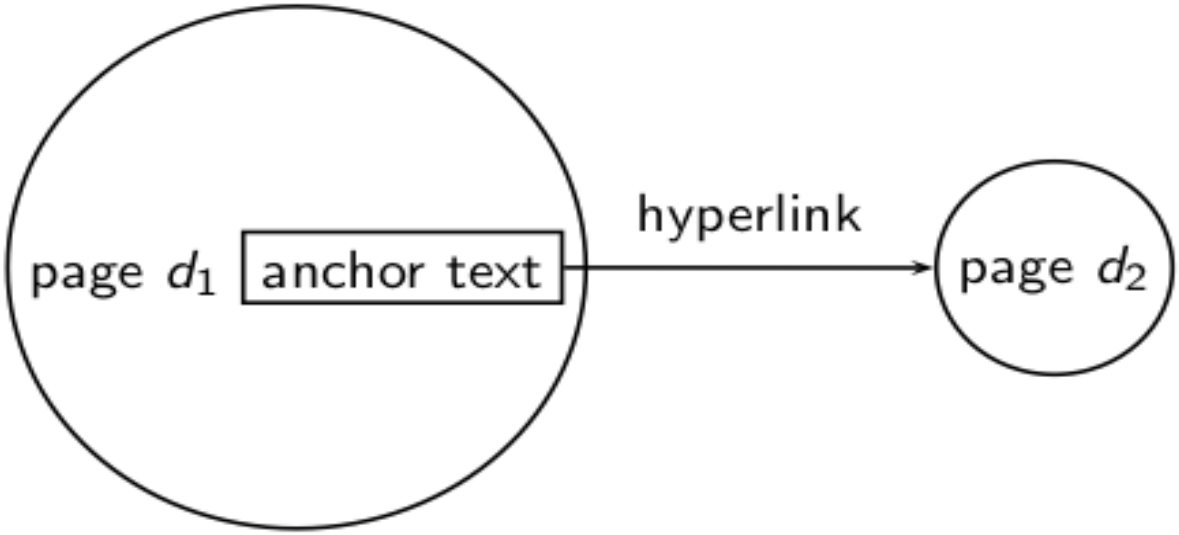
Κείμενο Άγκυρας



Anchor text (κείμενο άγκυρας) κείμενο που περιβάλει τον σύνδεσμο

- Παράδειγμα: “You can find cheap cars <a href =http://…> here . ”
- Παράδειγμα: “You can find <a href =http://…> cheap cars here . ”
- Anchor text: “You can find cheap cars here”

Σημασία των συνδέσεων



- 1η Υπόθεση: A hyperlink is a quality signal. (PageRank)
  - Η σύνδεση $d_1 \rightarrow d_2$ υποδηλώνει ότι ο συγγραφέας του d_1 θεωρεί το d_2 καλής ποιότητας και συναφές.
- 2η Υπόθεση: Το κείμενο της άγκυρας περιγράφει το περιεχόμενο του d_2.

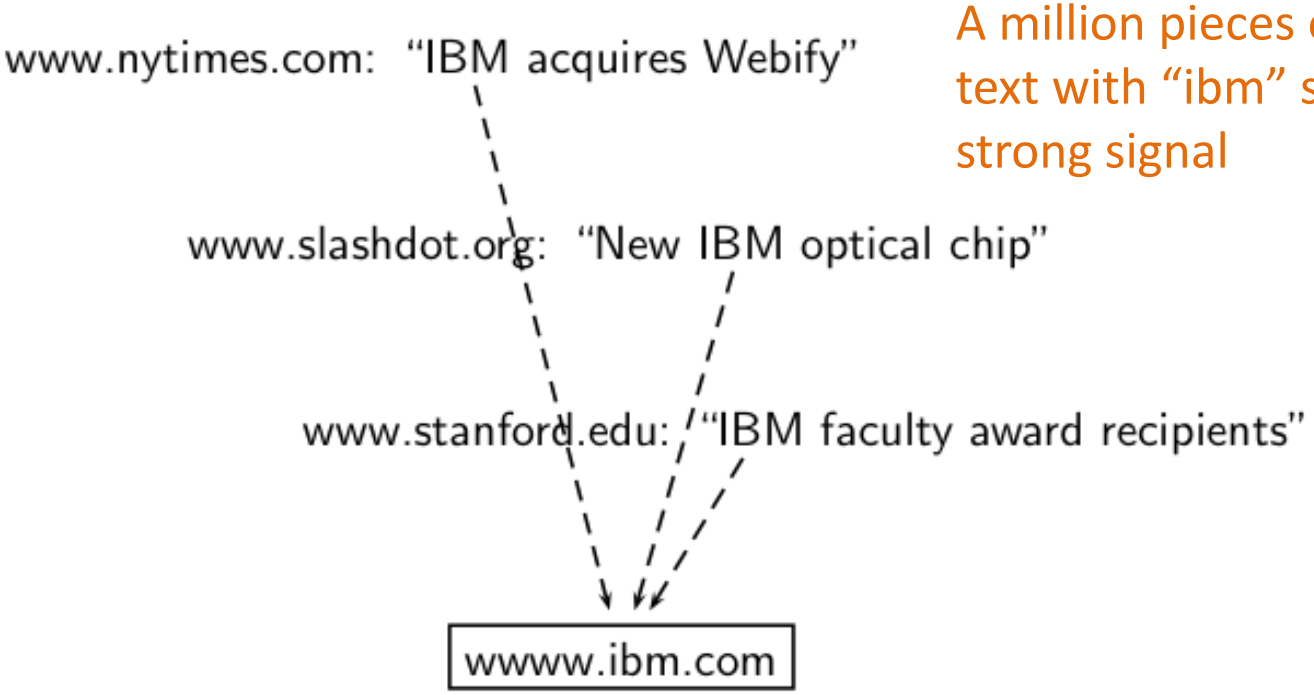
Κείμενο Άγκυρας

Χρήση μόνο **[text of d_2]** ή **[text of d_2] + [anchor text → d_2]**

- Αναζήτηση του [text of d_2] + [anchor text → d_2] συχνά πιο αποτελεσματική από την αναζήτηση μόνο του [text of d_2]
- Παράδειγμα: Ερώτημα *IBM*
  - Matches IBM's copyright page
  - Matches many spam pages
  - Matches IBM wikipedia article
  - *May not match IBM home page!* if IBM home page is mostly graphics

Κείμενο Άγκυρας

- Αναζήτηση με χρήση του [anchor text → d_2] καλύτερη για το ερώτημα IBM
  - Η σελίδα με τις περισσότερες εμφανίσεις του όρου *IBM* στο anchor text είναι η www.ibm.com

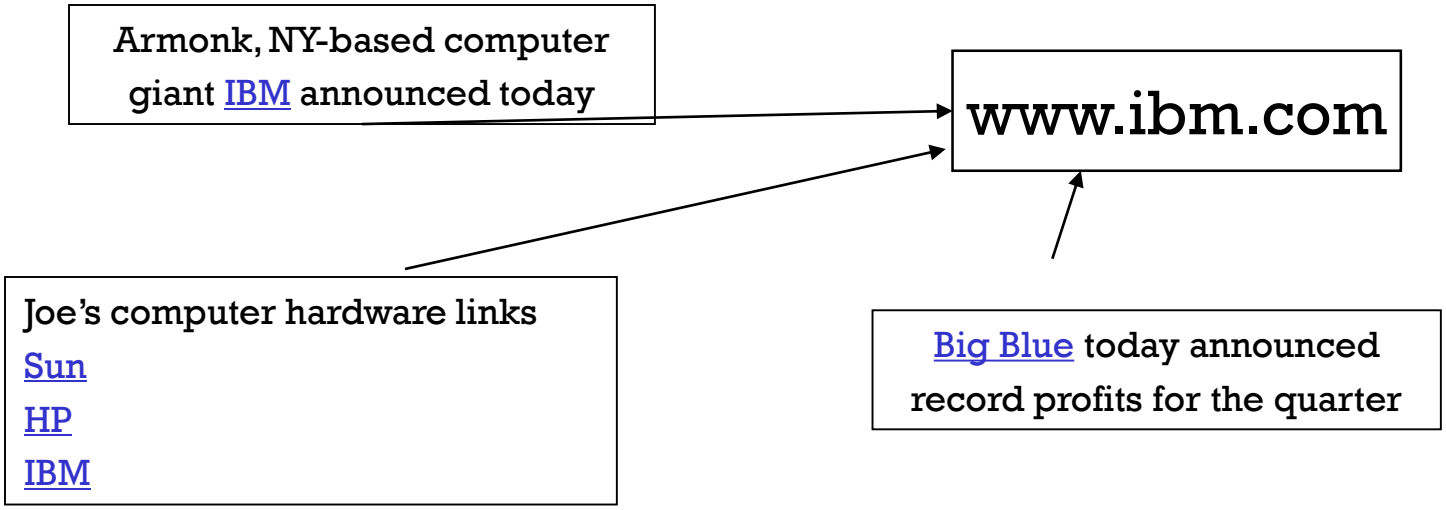


A million pieces of anchor text with "ibm" send a strong signal

Κείμενο Άγκυρας στο Ευρετήριο

Άρα: Το κείμενο στην άγκυρα αποτελεί καλύτερη περιγραφή του περιεχομένου της σελίδας από ό,τι το περιεχόμενο της

- Όταν κατασκευάζουμε το ευρετήριο για ένα έγγραφο *D*, συμπεριλαμβάνουμε (με κάποιο βάρος) και το κείμενο της άγκυρας των συνδέσεων που δείχνουν στο *D*.



- Χρήση: Χρήση idf για κοινούς όρους όπως Click, Here
- Επίσης, extended anchor text

Google Bombs

Google bomb: μια αναζήτηση με «κακά» αποτελέσματα εξαιτίας κακόβουλης χρήσης κειμένου άγκυρας

Η Google εισήγαγε μια νέα συνάρτηση βαρών το Ιανουάριο του 2007

- *Βάρος στο κείμενο άγκυρας ανάλογα με την εγκυρότητα/κύρος της σελίδα που το περιέχει*
  - Miserable failure (Bush 2004)
- Ακόμα κάποια υπόλοιπα: [dangerous cult] στο Google, Bing, Yahoo
  - the Church of Scientology
- Defused Google bombs: [dumb motherf…], [who is a failure?], [evil empire] [cheerful achievement]

Κείμενο Άγκυρας

- Άλλες εφαρμογές
  - Απόδοση βαρών/φιλτράρισμα στο γράφο
  - Για δημιουργία περιλήψεων μιας σελίδας

Υπόθεση 2: annotation of target

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)

Car rental Brasil

- Find a good hub
- Exploratory search “see what’s there”

Typing Queries: παραδείγματα

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

Τι ψάχνουν;

Δημοφιλή ερωτήματα

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

Και ανά χώρα

Τα ερωτήματα ακολουθούν επίσης power law κατανομή

Ανάγκες Χρηστών

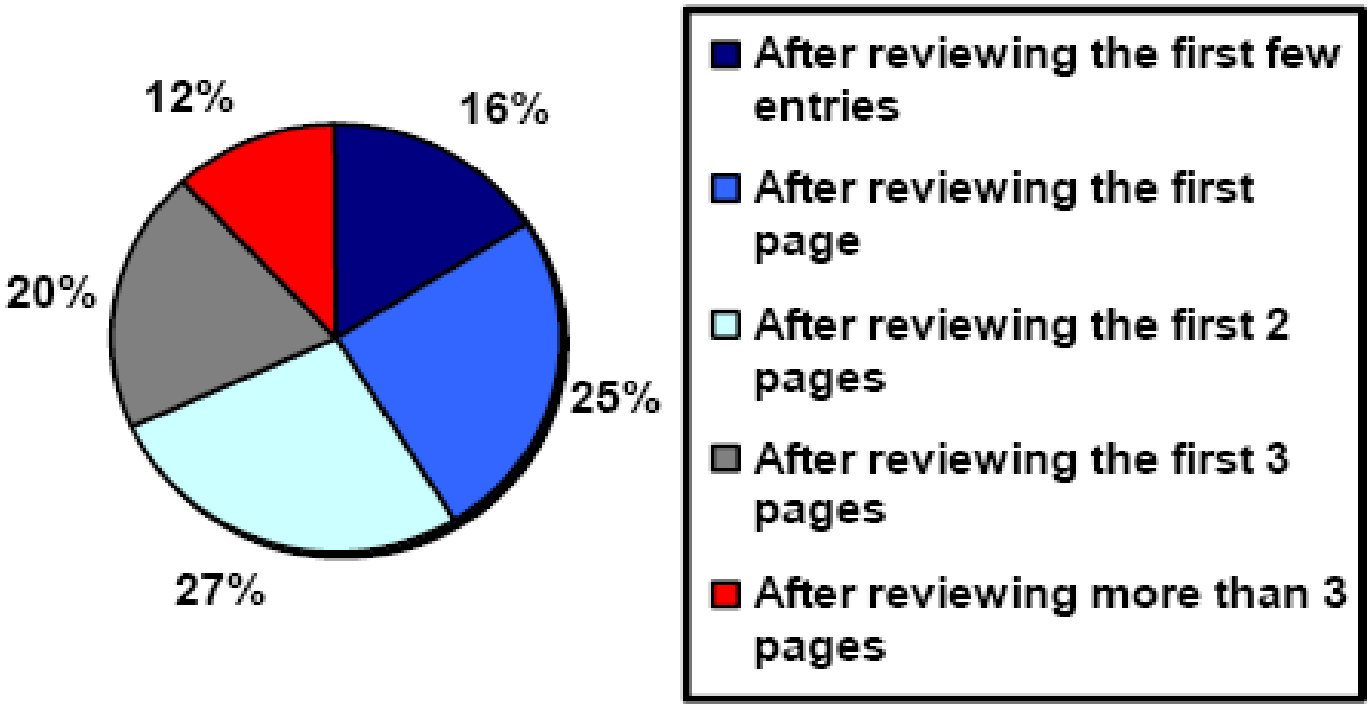
Επηρεάζει (ανάμεσα σε άλλα)

- 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
- Context of user in this session (e.g., previous query)
- Context provided by personal profile

Γεωγραφικό Περιεχόμενο

Τρεις σχετικές τοποθεσίες

1. Server (nytimes.com → New York)
2. Ιστοσελίδα (άρθρο των nytimes.com σχετικά με την Albania)
3. Χρήστης (βρίσκεται στο Palo Alto)

Εύρεση της τοποθεσίας του χρήστη

- IP address
- Πληροφορία που παρέχει ο χρήστης (πχ, στο user profile)
- Κινητό τηλέφωνπ

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

- ✓ δύσκολο NLP πρόβλημα

Χρήση context

Πως μπορούμε να χρησιμοποιήσουμε τα συμφραζόμενα (context) για να τροποποιήσουμε τα αποτελέσματα ενός ερωτήματος;

- Result restriction: Αγνοούμε τα ακατάλληλα αποτελέσματα
  - Στο χρήστη της google.fr δείξε μόνο .fr αποτελέσματα
- Ranking modulation: χρησιμοποίησέ μια γενική διάταξη και επανα-διέταξε τα αποτελέσματα με βάση τα συμφραζόμενα

Personalization/contextualization

Τι θα δούμε σήμερα

- Διαφημίσεις
- Θέματα Παραλληλίας
- Ταξινόμηση

ΔΙΑΦΗΜΙΣΕΙΣ

Διαφημίσεις

Graphical graph banners σε δημοφιλείς ιστοσελίδες (branding)

- ***cost per mil (CPM) model***: το κόστος προβολής του banner 1000 φορές (γνωστό και ως impressions)
- ***cost per click (CPC) model***: ο αριθμός των clicks στη διαφήμιση (που οδηγεί σε σελίδα από την οποία μπορεί να γίνει μια αγορά)
- *brand promotion* vs *transaction-oriented advertising*

Ιστορία

- Αρχικές μηχανές βασισμένες σε λέξεις κλειδιά γύρω στο 1995-1997
  - Altavista, Excite, Infoseek, Inktomi, Lycos
- Paid search διάταξη: Goto (μετεξέλιξη σε Overture.com → Yahoo!)
  - Το αποτέλεσμα της αναζήτησης εξαρτάται από το πόσο πλήρωσες
  - Δημοπρασία (Auction) για λέξεις κλειδιά: η λέξη ***casino*** ακριβή!

Διαφημίσεις στο Goto

Ως αποτέλεσμα μια ερώτησης *q*, η Goto

- επιστρέφει *τις σελίδες των διαφημιζόμενων* που πόνταραν (bid) για το *q*, σε διάταξη με βάση την προσφορά τους.
- Όταν ο χρήστες επέλεγε (clicked) σε ένα από τα αποτελέσματα, ο αντίστοιχος διαφημιζόμενος πλήρωνε τη Goto
  - Αρχικά, η πληρωμή ίση με την προσφορά bid για το q

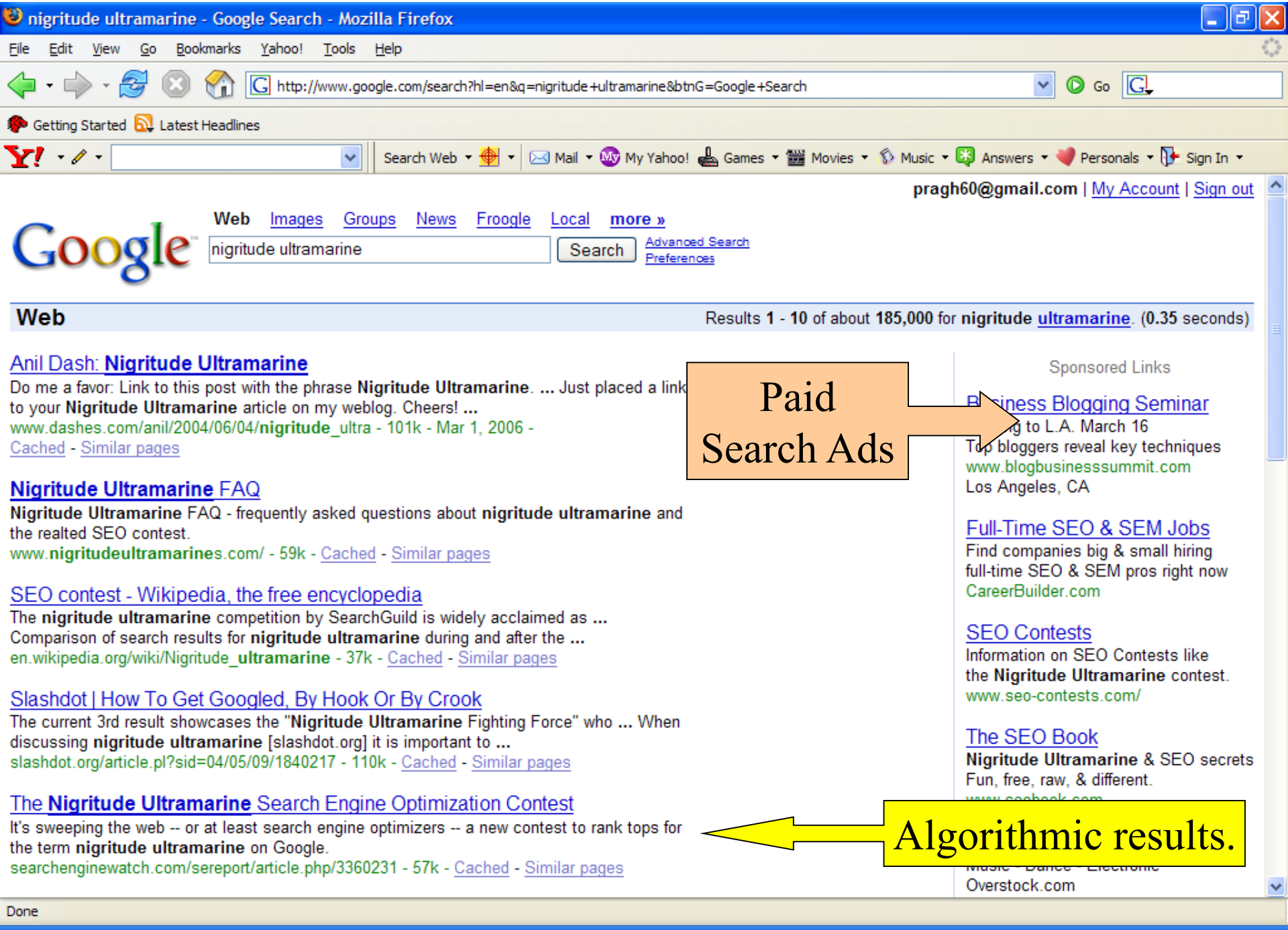
Sponsored search or *Search advertising*

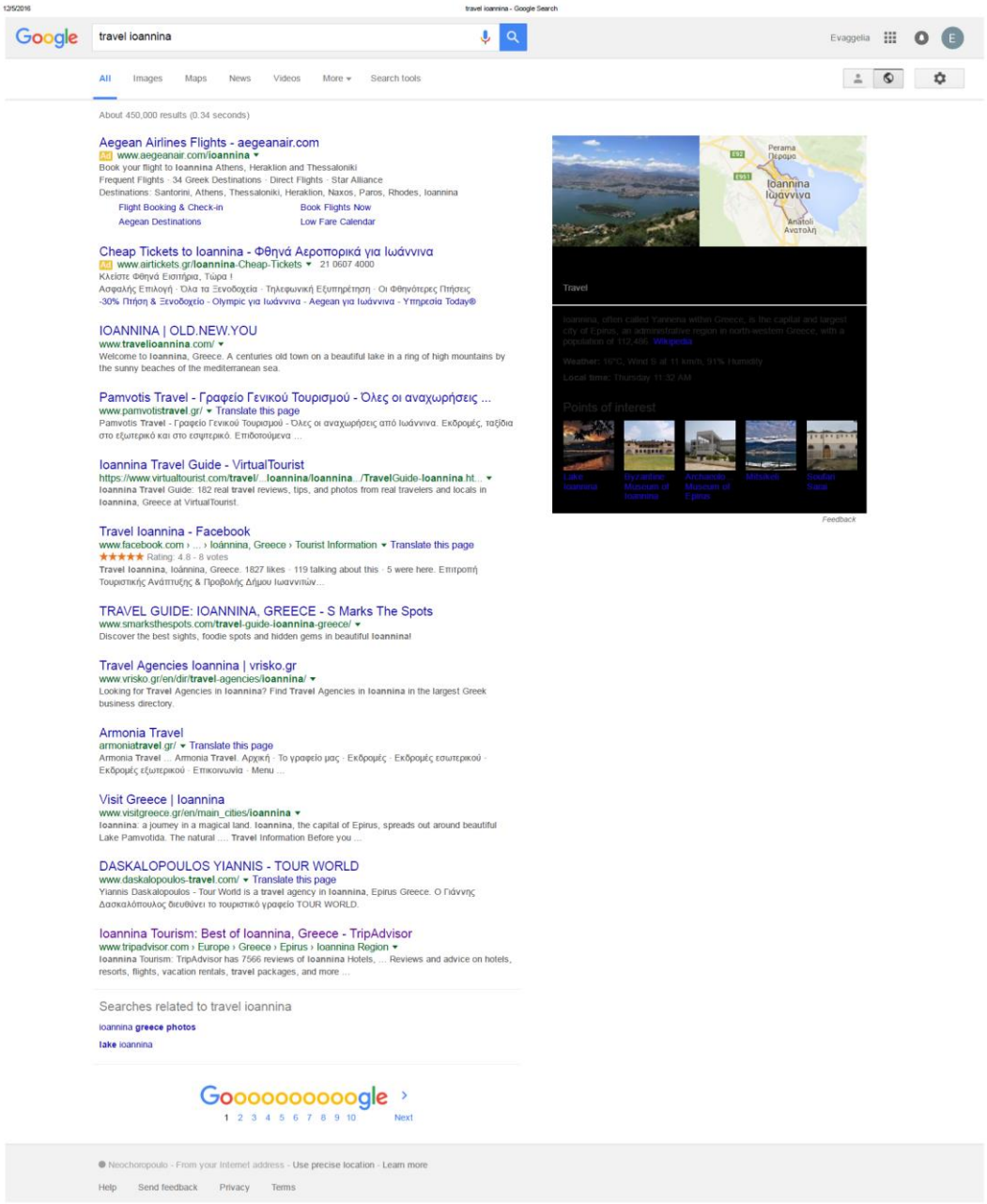
Διαφημίσεις στο Goto

Διαφημίσεις

Συνήθως παρέχονται

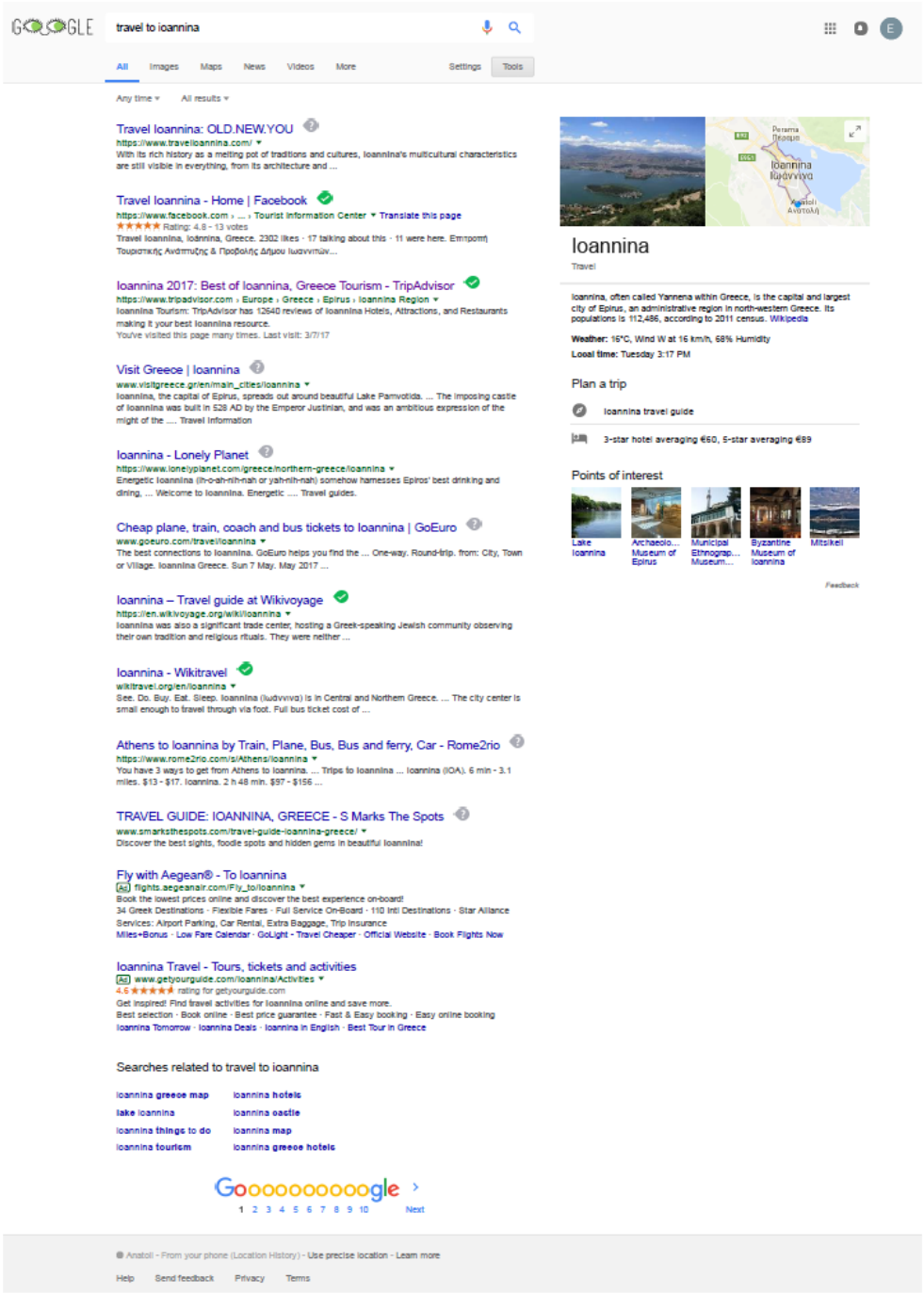
- ***pure search results*** (γνωστά και ως αλγοριθμικά ή *organic search* αποτελέσματα) ως βασική απάντηση στο ερώτημα του χρήστη,
- μαζί με ***sponsored search results*** τα οποία παρουσιάζονται ξεχωριστά και διακριτά στα δεξιά των αλγοριθμικών αποτελεσμάτων


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
Go
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

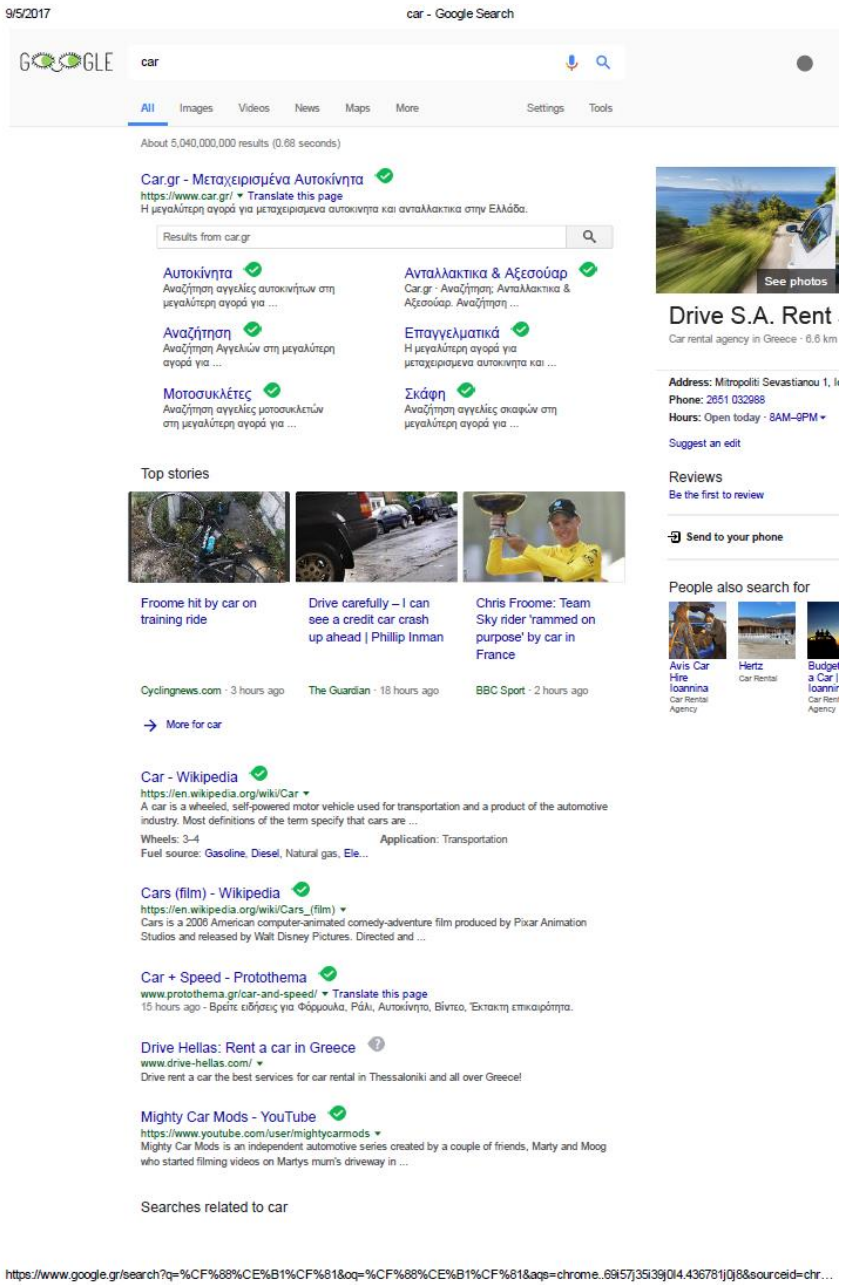


2016

location


travel to ioannina
All
Images
Maps
News
Videos
More
Settings
Tools
Any time
All results
Travel Ioannina: OLD.NEW.YOU
https://www.travelioannina.com/
With its rich history as a melting pot of traditions and cultures, Ioannina's multicultural characteristics are still visible in everything, from its architecture and ...
Travel Ioannina - Home | Facebook
https://www.facebook.com › ... › Tourist Information Center
Translate this page
Rating: 4.8 - 13 votes
Travel Ioannina, Ioánnina, Greece. 2302 likes · 17 talking about this · 11 were here. Επίσημη Τουριστικής Ανάπτυξης & Προβολής Δήμου Ιωαννιτών...
Ioannina 2017: Best of Ioannina, Greece Tourism - TripAdvisor
https://www.tripadvisor.com › Europe › Greece › Epirus › Ioannina Region
Ioannina Tourism: TripAdvisor has 12640 reviews of Ioannina Hotels, Attractions, and Restaurants making it your best Ioannina resource.
You've visited this page many times. Last visit: 3/7/17
Visit Greece | Ioannina
www.visitgreece.gr/en/main_cities/ioannina
Ioannina, the capital of Epirus, spreads out around beautiful Lake Pamvotida. ... The imposing castle of Ioannina was built in 528 AD by the Emperor Justinian, and was an ambitious expression of the might of the .... Travel Information
Ioannina - Lonely Planet
https://www.lonelyplanet.com/greece/northern-greece/ioannina
Energetic Ioannina (ih-o-ah-nih-nah or yah-nih-nah) somehow harnesses Epiros' best drinking and dining, ... Welcome to Ioannina. Energetic .... Travel guides.
Cheap plane, train, coach and bus tickets to Ioannina | GoEuro
www.goeuro.com/travel/ioannina
The best connections to Ioannina. GoEuro helps you find the ... One-way. Round-trip. from: City, Town or Village. Ioannina Greece. Sun 7 May. May 2017 ...
Ioannina – Travel guide at Wikivoyage
https://en.wikivoyage.org/wiki/Ioannina
Ioannina was also a significant trade center, hosting a Greek-speaking Jewish community observing their own tradition and religious rituals. They were neither ...
Ioannina - Wikitravel
wikitravel.org/en/Ioannina
See. Do. Buy. Eat. Sleep. Ioannina (Ιωάννινα) is in Central and Northern Greece. ... The city center is small enough to travel through via foot. Full bus ticket cost of ...
Athens to Ioannina by Train, Plane, Bus, Bus and ferry, Car - Rome2rio
https://www.rome2rio.com/s/Athens/Ioannina
You have 3 ways to get from Athens to Ioannina. ... Trips to Ioannina ... Ioannina (IOA). 6 min - 3.1 miles. $13 - $17. Ioannina. 2 h 48 min. $97 - $156 ...
TRAVEL GUIDE: IOANNINA, GREECE - S Marks The Spots
www.smarksthespots.com/travel-guide-ioannina-greece/
Discover the best sights, foodie spots and hidden gems in beautiful Ioannina!
Fly with Aegean® - To Ioannina
Ad flights.aegeanair.com/Fly_to/Ioannina
Book the lowest prices online and discover the best experience on-board!
34 Greek Destinations · Flexible Fares · Full Service On-Board · 110 Intl Destinations · Star Alliance
Services: Airport Parking, Car Rental, Extra Baggage, Trip Insurance
Miles+Bonus · Low Fare Calendar · GoLight · Travel Cheaper · Official Website · Book Flights Now
Ioannina Travel - Tours, tickets and activities
Ad www.getyourguide.com/Ioannina/Activities
4.6 rating for getyourguide.com
Get inspired! Find travel activities for Ioannina online and save more.
Best selection · Book online · Best price guarantee · Fast & Easy booking · Easy online booking
Ioannina Tomorrow · Ioannina Deals · Ioannina in English · Best Tour in Greece
Searches related to travel to ioannina
ioannina greece map
ioannina hotels
lake ioannina
ioannina castle
ioannina things to do
ioannina map
ioannina tourism
ioannina greece hotels
Goooooooooogle
1 2 3 4 5 6 7 8 9 10 Next
Anatoli - From your phone (Location History) - Use precise location - Learn more
Help
Send feedback
Privacy
Terms
Ioannina
Travel
Ioannina, often called Yannena within Greece, is the capital and largest city of Epirus, an administrative region in north-western Greece. Its populations is 112,486, according to 2011 census. Wikipedia
Weather: 16°C, Wind W at 16 km/h, 68% Humidity
Local time: Tuesday 3:17 PM
Plan a trip
Ioannina travel guide
3-star hotel averaging €60, 5-star averaging €89
Points of interest
Lake Ioannina
Archaeolo... Museum of Epirus
Municipal Ethnograp... Museum...
Byzantine Museum of Ioannina
Mitsikeli
Feedback

2017


9/5/2017
car - Google Search
GOOGLE
car
All
Images
Videos
News
Maps
More
Settings
Tools
About 5,040,000,000 results (0.68 seconds)
Car.gr - Μεταχειρισμένα Αυτοκίνητα
https://www.car.gr/ ▾ Translate this page
Η μεγαλύτερη αγορά για μεταχειρισμένα αυτοκίνητα και ανταλλακτικά στην Ελλάδα.
Results from car.gr
Αυτοκίνητα
Αναζήτηση αγγελίες αυτοκινήτων στη μεγαλύτερη αγορά για ...
Ανταλλακτικά & Αξεσούαρ
Car.gr · Αναζήτηση: Ανταλλακτικα & Αξεσούαρ. Αναζήτηση ...
Αναζήτηση
Αναζήτηση Αγγελιών στη μεγαλύτερη αγορά για ...
Επαγγελματικά
Η μεγαλύτερη αγορά για μεταχειρισμένα αυτοκίνητα και ...
Μοτοσυκλέτες
Αναζήτηση αγγελίες μοτοσυκλετών στη μεγαλύτερη αγορά για ...
Σκάφη
Αναζήτηση αγγελίες σκαφών στη μεγαλύτερη αγορά για ...
See photos
Drive S.A. Rent
Car rental agency in Greece · 6.6 km
Address: Mitropoliti Sevastianou 1,
Phone: 2651 032988
Hours: Open today · 8AM–9PM
Suggest an edit
Reviews
Be the first to review
Send to your phone
People also search for
Avis Car Hire Ioannina
Car Rental Agency
Hertz
Car Rental
Top stories
Froome hit by car on training ride
Cyclingnews.com · 3 hours ago
Drive carefully – I can see a credit car crash up ahead | Phillip Inman
The Guardian · 18 hours ago
Chris Froome: Team Sky rider 'rammed on purpose' by car in France
BBC Sport · 2 hours ago
More for car
Car - Wikipedia
https://en.wikipedia.org/wiki/Car ▾
A car is a wheeled, self-powered motor vehicle used for transportation and a product of the automotive industry. Most definitions of the term specify that cars are ...
Wheels: 3–4
Application: Transportation
Fuel source: Gasoline, Diesel, Natural gas, Ele...
Cars (film) - Wikipedia
https://en.wikipedia.org/wiki/Cars_(film) ▾
Cars is a 2006 American computer-animated comedy-adventure film produced by Pixar Animation Studios and released by Walt Disney Pictures. Directed and ...
Car + Speed - Protothema
www.protothema.gr/car-and-speed/ ▾ Translate this page
15 hours ago - Βρείτε ειδήσεις για Φόρμουλα, Ράλι, Αυτοκίνητο, Βίντεο, Έκτακτη επικαιρότητα.
Drive Hellas: Rent a car in Greece
www.drive-hellas.com/ ▾
Drive rent a car the best services for car rental in Thessaloniki and all over Greece!
Mighty Car Mods - YouTube
https://www.youtube.com/user/mightycarmods ▾
Mighty Car Mods is an independent automotive series created by a couple of friends, Marty and Moog who started filming videos on Martys mum's driveway in ...
Searches related to car
https://www.google.gr/search?q=%CF%88%CE%B1%CF%81&oq=%CF%88%CE%B1%CF%81&aqs=chrome..69i57j35i39j0l4.436781j0j8&sourceid=chr...

2017

Διαφημίσεις

- **Search Engine Marketing (SEM)**

Κλάδος της διαφήμισης: κατανόηση του πως γίνεται η διάταξη των αποτελεσμάτων και ποιο ποσό της διαφημιστικής καμπάνιας να δοθεί σε ποιες λέξεις και σε ποιες μηχανές

Click spam: clicks σε sponsored search results από μη πραγματικούς χρήστες

πχ, από έναν αντίπαλο διαφημιστή

Διαφημίσεις

Paid inclusion: πληρωμή για να συμπεριληφθεί μια σελίδα στο ευρετήριο της μηχανής αναζήτησης

Διαφορετικές μηχανές ακολουθούν διαφορετικές πολιτικές σχετικά με το αν επιτρέπουν *paid inclusion* και αν αυτό επηρεάζει τη διάταξη των αποτελεσμάτων της αναζήτησης.

Παρόμοια προβλήματα με τηλεόραση/εφημερίδες

Διάταξη διαφημίσεων

- Οι διαφημιστές κάνουν *προσφορές για λέξεις κλειδιά* σε δημοπρασίες
- Ανοικτό σύστημα: Οποιοσδήποτε μπορεί να συμμετάσχει και να κάνει προσφορές
- Οι διαφημιστές χρεώνονται μόνο όταν κάποιος επιλέγει (clicks) στη διαφήμισή τους
- Σημαντική περιοχή για τις μηχανές αναζήτησης – *computational advertising.*
  - μια μικρή αύξηση της χρέωσης σε κάθε διαφήμιση μπορεί να οδηγήσει σε δις επιπρόσθετο κέρδος

Διάταξη διαφημίσεων

Πως αποφασίζεται μέσω της δημοπρασίας η θέση της διαφήμισης και η τιμή

- Βασίζεται στο 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 του διαφημιστή για ένα click
- CTR: click-through rate: το ποσοστό των χρηστών που επιλέγουν (click on) μια διαφήμιση όταν η διαφήμιση προβάλλεται
  - το CTR είναι μια μέτρηση της συνάφειας
- ad rank: bid × CTR: εξισορρόπηση ανάμεσα (i) σε πόσα χρήματα είναι διατεθειμένος κάποιος διαφημιστής να πληρώσει, και (ii) στο πόσο σχετική είναι η διαφήμιση
- rank: θέση στη διάταξη της δημοπρασίας
- paid: second price auction price που πληρώνει ο διαφημιστής

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: Ο διαφημιστής πληρώνει το μικρότερο ποσό ώστε να διατηρήσει τη θέση του στη δημοπρασία (συν 1 cent).

$price_1 \times CTR_1 = bid_2 \times CTR_2$ (αυτό έχει ως αποτέλεσμα $rank_1=rank_2$)

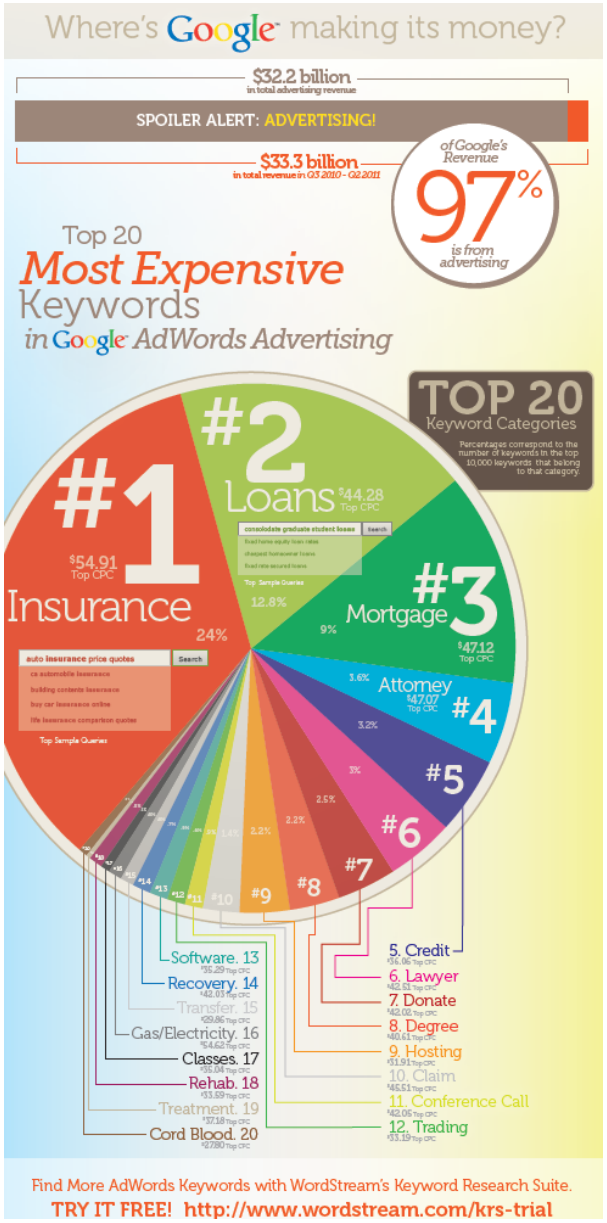
$price_1 = bid_2 \times CTR_2 / CTR_1$

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

Λέξεις κλειδιά με τα μεγαλύτερα bids

από το 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
$54.91
Top CPC
Insurance
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

2016

Search ads: A win-win-win?

- Η μηχανή αναζήτησης κερδίζει χρήματα κάθε φορά που κάποιος επιλέγει (clicks) μια διαφήμιση
- Ο χρήστης επιλέγει μια διαφήμιση μόνο αν τον ενδιαφέρει
  - Οι μηχανές τιμωρούν μη συναφές ή παραπλανητικό περιεχόμενο
  - Ως αποτέλεσμα οι χρήστες συνήθως είναι ικανοποιημένοι από το τι βρίσκουν όταν επιλέγουν μια διαφήμιση
- Ο διαφημιστής βρίσκει νέους πελάτες με ένα αποδοτικό από άποψη κόστους τρόπο.

Not a win-win-win: Keyword arbitrage

- Αγόρασε μια λέξη κλειδί στο Google
- Μετά επανα-προώθησε την κυκλοφορία σε κάποιον τρίτο που πληρώνει περισσότερα (ένα φτηνό keyword σε ένα ακριβό)
  - E.g., μια σελίδα γεμάτη διαφημίσεις
- Ad spammers keep inventing new tricks.
- The search engines need time to catch up with them.

62

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.

Google search options

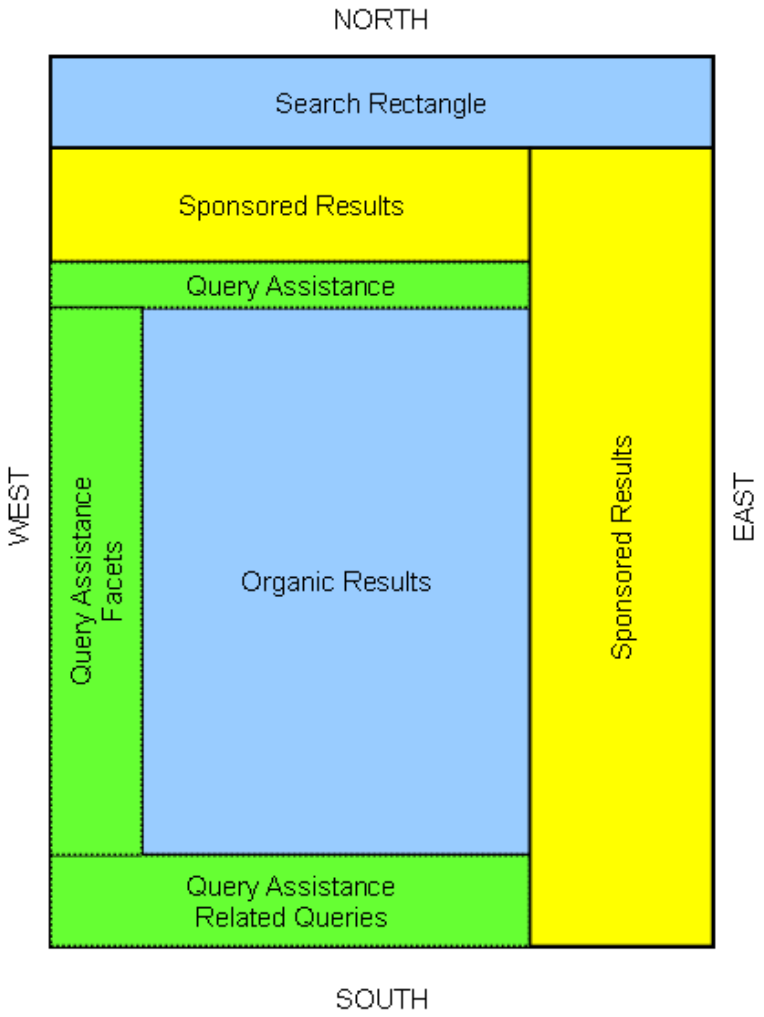
Settings: Advanced search

Turn on safe search

Hide private results

History

SERP Layout


NORTH
Search Rectangle
Sponsored Results
Query Assistance
Query Assistance Facets
Organic Results
Sponsored Results
WEST
EAST
Query Assistance Related Queries
SOUTH

ΚΡΙΤΙΚΗ

Ranking, auto-completion, …

- Search engines
- Facebook news feed
- ..

- Fairness
- Transparency
- Accountability
- Eco-chambers, information bubbles

Θα τελειώσουμε με μερικά αρνητικά παραδείγματα ..

Case Study: Gender bias in image search [CHI15]

What images do people choose to represent careers?

In search results:

- evidence for *stereotype exaggeration*
- systematic *underrepresentation of women*

- People rate search results *higher* when they are *consistent* with stereotypes for a career
- Shifting the representation of gender in image search results can *shift* *people's perceptions* about real-world distributions. (after search slight increase in their believes)

Tradeoff between high-quality result and broader societal goals for equality of representation

Case Study: Latanya

The importance of being Latanya

Names used predominantly by *black men and women* are much more likely to generate *ads* related *to arrest records*, than names used predominantly by white men and women.

Case Study: AdFisher

Tool to automate the creation of *behavioral* and *demographic* profiles.

http://possibility.cylab.cmu.edu/adfisher/

- setting gender = female results in less ads for high-paying jobs
- browsing substance abuse websites leads to rehab ads

Fairness: google search and autocomplete

Donald Tramp accused Google “suppressing negative information” about Clinton
Autocomplete feature - “hillary clinton cri” vs “donald tramp cri”

Autocomplete:

- are jews
- are women

https://www.theguardian.com/us-news/2016/sep/29/donald-trump-attacks-biased-lester-holt-and-accuses-google-of-conspiracy

https://www.theguardian.com/technology/2016/dec/04/google-democracy-truth-internet-search-facebook?CMP=fb_gu

Google+ names

Google+ tries to classify Real vs Fake names

Fairness problem:

– Most training examples standard white American names

– Ethnic names often unique, much fewer training examples

Likely outcome:

Prediction accuracy *worse on ethnic names*

Katya Casio. *"Due to Google's ethnocentricity I was prevented from using my real last name (my nationality is: Tungus and Sami)"*
Google Product Forums

Other

LinkedIn: female vs male names (for female prompts suggestions for male, e.g., Andrea Jones" to "Andrew Jones," Danielle to Daniel, Michaela to Michael and Alexa to Alex.)
http://www.seattletimes.com/business/microsoft/how-linkedins-search-engine-may-reflect-a-bias/

Flickr: auto-tagging system labels images of black people as apes or animals and concentration camps as sport or jungle jyms.
https://www.theguardian.com/technology/2015/may/20/flickr-complaints-offensive-auto-tagging-photos

Airbnb: race discrimination
Against guest
http://www.debiasyourself.org/
Community commitment
http://blog.airbnb.com/the-airbnb-community-commitment/
Non-black hosts can charge ~12% more than black hosts
Edelman, Benjamin G. and Luca, Michael, Digital Discrimination: The Case of Airbnb.com (January 10, 2014). Harvard Business School NOM Unit Working Paper No. 14-054.

Google maps: China is about 21% larger by pixels when shown in Google Maps for China
Gary Soeller, Karrie Karahalios, Christian Sandvig, and Christo Wilson: MapWatch: Detecting and Monitoring International Border Personalization on Online Maps. Proc. of WWW. Montreal, Quebec, Canada, April 2016

Some tools regarding information bubbles

Is your news feed a bubble?
PolitEcho shows you the political biases of your Facebook friends and news feed.
http://politecho.org/

Step into someone else's Twitter feed
Feeds belong to users classified as having left or right-leaning political ideologies.
https://flipfeed.media.mit.edu/

Insert a clearly-marked article into your Facebook Feed reflecting who you are trying to understand better.
https://www.escapeyourbubble.com/

Μια ματιά στην ΑΠ πολύ μεγάλης κλίμακας

Μερικοί αριθμοί

- The Indexed Web contains at least 1.71 billion pages (Sunday, 16 March, 2014).

- Each year, Google changes its search algorithm around 500–600 times

http://moz.com/google-algorithm-change

Κατανεμημένα ΣΑΠ

- Για ευρετήριο κλίμακας web
  Χρήση κατανεμημένου cluster
- Επειδή μια μηχανή είναι επιρρεπής σε αποτυχία
  – Μπορεί απροσδόκητα να γίνει αργή ή να αποτύχει
- Χρησιμοποίηση πολλών μηχανών

Web search engine data centers

- Οι μηχανές αναζήτησης χρησιμοποιούν data centers (Google, Bing, Baidu) κυρίως από commodity μηχανές. *Γιατί; (fault tolerance)*
- Τα κέντρα είναι διάσπαρτα σε όλο τον κόσμο.
- Εκτίμηση: Google ~1 million servers, 3 million processors/cores (Gartner 2007)

https://www.google.com/about/datacenters/

Λίγα εισαγωγικά για το MapReduce και τη χρήση του στην κατασκευή του ευρετηρίου

Κατανομή των Ευρετηρίων

How to distribute the term index across a large computer cluster that supports querying.

Two alternatives index implementations

- *partitioning by terms* or *global* index organization, and
- *partitioning by documents* or *local* index organization.

Κατανομή βάσει Όρων

- Index terms partitioned into subsets,
- Each subset resides at a node.
- Along with the terms at a node, we keep their postings

A query is routed to the nodes corresponding to its query terms.

In principle, this allows greater concurrency since a stream of queries with different query terms would hit different sets of machines.

Κατανομή βάσει Εγγράφων

- Documents partitioned into subsets
- Each subset resides in a node
- Each node contains the index for a subset of all documents.

A query is distributed to all nodes, with the results from various nodes being merged before presentation to the user.

Κατανομή βάσει Όρων

- In principle, index partition allows *greater concurrency*, since a stream of queries with different query terms would hit different sets of machines.

- In practice, partitioning indexes by vocabulary terms turns out to be non-trivial.

Κατανομή βάσει Όρων

- Multi-word queries require the *sending of long postings lists* between sets of nodes for merging, and the cost of this can outweigh the greater concurrency.
- *Load balancing* the partition is governed not by an a priori analysis of relative term frequencies, but rather by the *distribution of query terms and their co-occurrences*, which can drift with time or exhibit sudden bursts.
- More *difficult implementation*.

Κατανομή βάσει Εγγράφων

More common

- trades more local disk seeks for less inter-node communication.
- One difficulty: *global statistics* used in scoring - such as idf
  - must be computed across the entire document collection even though the index at any single node only contains a subset of the documents.
  - Computed by distributed ``background'' processes that periodically refresh the node indexes with fresh global statistics.

Μέθοδος Κατανομής Εγγράφων

How to distributed documents to nodes?

- Hash of each URL to nodes

At query time,

the query is broadcast to each of the nodes, each node sends each top k results which are merged to find the top k documents for the query

Μέθοδος Κατανομής Εγγράφων

A common implementation heuristic:

Partition the document collection into

- indexes of documents that are more likely to score highly on most queries and
- low-scoring indexes with the remaining documents

Only search the low-scoring indexes when there are too few matches in the high-scoring indexes

Κατανεμημένα ΣΑΠ

1. *Term-partitioned:* one machine handles a subrange of terms
2. *Document-partitioned:* one machine handles a subrange of documents

- most search engines use a document-partitioned index, better load balancing, etc.

Google index

- index partitioned by document IDs into pieces called shards
- each shard is replicated onto multiple servers
- initially, from hard disk drives, now enough servers to keep a copy of the whole index in main memory

Google database Spanner (NewSQL) – (μέρος της πλατφόρμας αλλά και του cloud) F1

Παράλληλη κατασκευή

- Maintain a *master* machine directing the indexing job – considered “safe”.
- Break up indexing into sets of (parallel) tasks.

- Master machine assigns each task to an idle machine from a pool.

Parallel tasks

- Break the input document collection into *splits*
- Each split is a subset of documents (corresponding to blocks in BSBI/SPIMI)

- We will use two sets of parallel tasks
  - Parsers
  - Inverters

Parsers

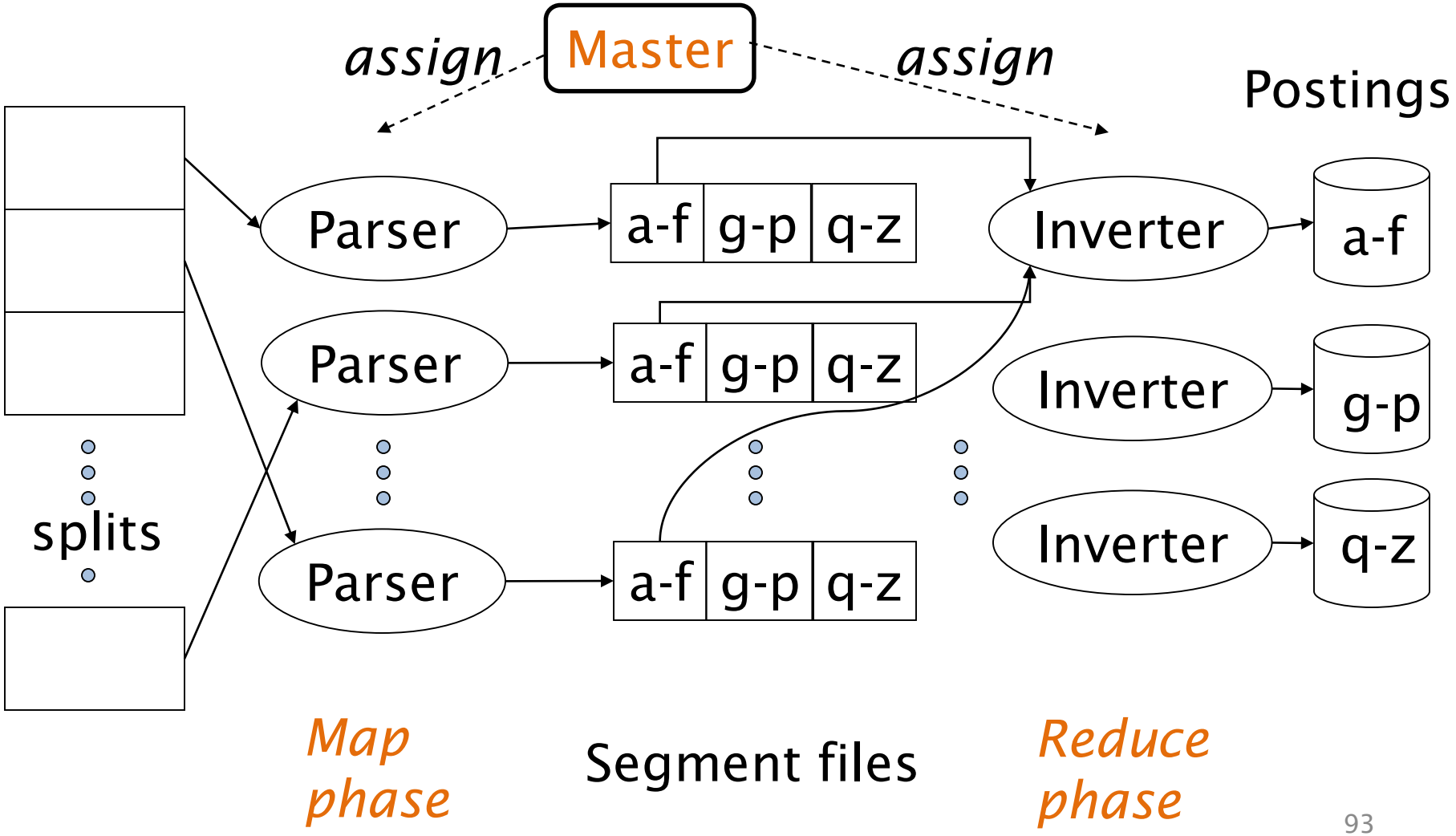
- Master assigns a split to an idle parser machine
- Parser reads a document at a time and emits (term, doc) pairs
- Parser writes pairs into *j* partitions
  - Each partition is for a range of terms' first letters

  (e.g., ***a-f, g-p, q-z***) – here *j* = 3.

Inverters

- An inverter collects all (term, doc) pairs (= postings) for one term-partition.
- Sorts and writes to postings lists

Παράλληλη κατασκευή



MapReduce

- The index construction algorithm we just described is an instance of *MapReduce*.
- MapReduce (Dean and Ghemawat 2004) is a robust and conceptually simple framework for distributed computing without having to write code for the distribution part.
- They describe the Google indexing system (ca. 2002) as consisting of a number of phases, each implemented in MapReduce.

*open source implementation as part of Hadoop**

**http://hadoop.apache.org/*

Παράδειγμα κατασκευής ευρετηρίου σε MapReduce

Το γενικό σχήμα των συναρτήσεων map και reduce

- map: input → list(***key***, value)
- reduce: (***key***, list(value)) → output

Εφαρμογή στην περίπτωση της κατασκευής ευρετηρίου

- map: collection → list(termID, docID)
- reduce: (<termID1, list(docID)>, <termID2, list(docID)>, …) → (postings list1, postings list2, …)

Example for index construction

Map:

- d1 : C came, C c'ed.
- d2 : C died. →

<C,d1>, <came,d1>, <C,d1>, <c'ed, d1>, <C, d2>, <died,d2>

Reduce:

- (<C,(d1,d2,d1)>, <died,(d2)>, <came,(d1)>, <c'ed,(d1)>) →

(<C,(d1:2,d2:1)>, <died,(d2:1)>, <came,(d1:1)>, <c'ed,(d1:1)>)

MapReduce

- Index construction was just one phase.
- Another phase: transforming *a term-partitioned index* into a *document-partitioned index.*

- most search engines use a document-partitioned index

How search (in Google) works

Knowledge graph

https://www.google.com/search/howsearchworks/

Many features to combine for ranking
classification or learning to rank model

Using Supervised Learning

Given a collection of records (*training set*)

Each record contains
a set of *attributes (features) +* the *class attribute*.

Find a *model* for the class attribute as a function of the values of other attributes.

Goal: previously unseen records should be assigned a class as accurately as possible.

A test set is used to determine the accuracy of the model.

Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

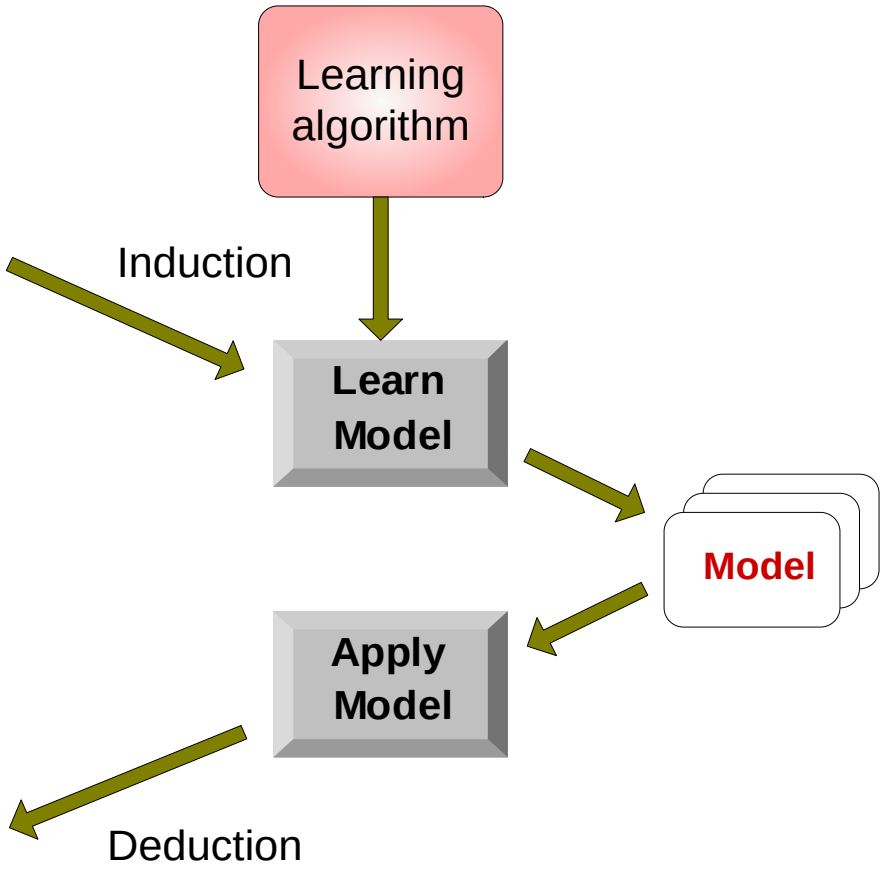
Illustrating the Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



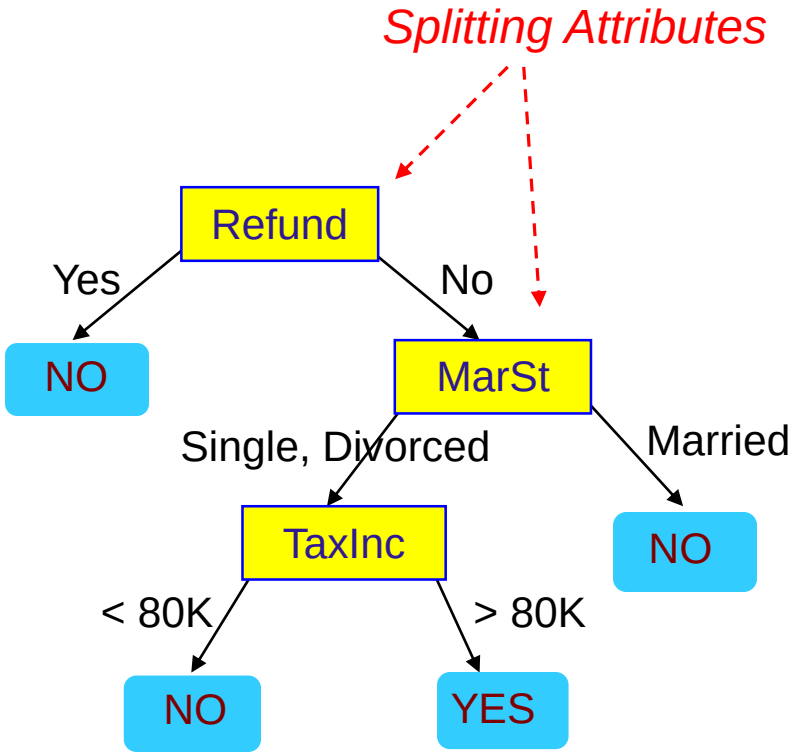
Classification Techniques

- Decision Tree based Methods
- Rule-based Methods
- Memory based reasoning
- Neural Networks
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines
- Logistic Regression

Example of a Decision Tree

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



Model: Decision Tree

Why is ML needed now?

- Modern systems – especially on the Web – use a great number of features:
  - Arbitrary useful features – not a single unified model
  - Log frequency of query word in anchor text?
  - Query word in color on page?
  - # of images on page?
  - # of (out) links on page?
  - PageRank of page?
  - URL length?
  - URL contains "~"?
  - Page edit recency?
  - Page length?
- The *New York Times* (2008-06-03) quoted Amit Singhal as saying Google was using over 200 such features.

Classification problem

Web graph (e.g., PageRank)
Document statistics
Document classifier (e.g., spam)
Query features
Term features
Proximity features
Location features
Time features
Clicks
Topical matching

Απλό παράδειγμα

- Collect a training corpus of (*q, d, r*) triples
  - Relevance *r* is here binary (but may be multiclass, with 3–7 values)
  - Document is represented by a feature vector

    x = (α, ω) α is cosine similarity, ω is minimum query window size
    - ω is the the shortest text span that includes all query words
    - Query term proximity is a **very important** new weighting factor
  - Train a machine learning model to predict the class *r* of a document-query pair

example	docID	query	cosine score	ω	judgment
Φ_1	37	linux operating system	0.032	3	*relevant*
Φ_2	37	penguin logo	0.02	4	*nonrelevant*
Φ_3	238	operating system	0.043	2	*relevant*
Φ_4	238	runtime environment	0.004	2	*nonrelevant*
Φ_5	1741	kernel layer	0.022	3	*relevant*
Φ_6	2094	device driver	0.03	2	*relevant*
Φ_7	3191	device driver	0.027	5	*nonrelevant*

"Learning to rank"

- Classification probably isn't the right way to think about approaching ad hoc IR:
  - Classification problems: Map to a unordered set of classes
  - Regression problems: Map to a real value
  - Ordinal regression problems: Map to an *ordered* set of classes

- This formulation gives extra power:
  - Relations between relevance levels are modeled
  - Documents are good versus other documents for query given collection; not an absolute scale of goodness

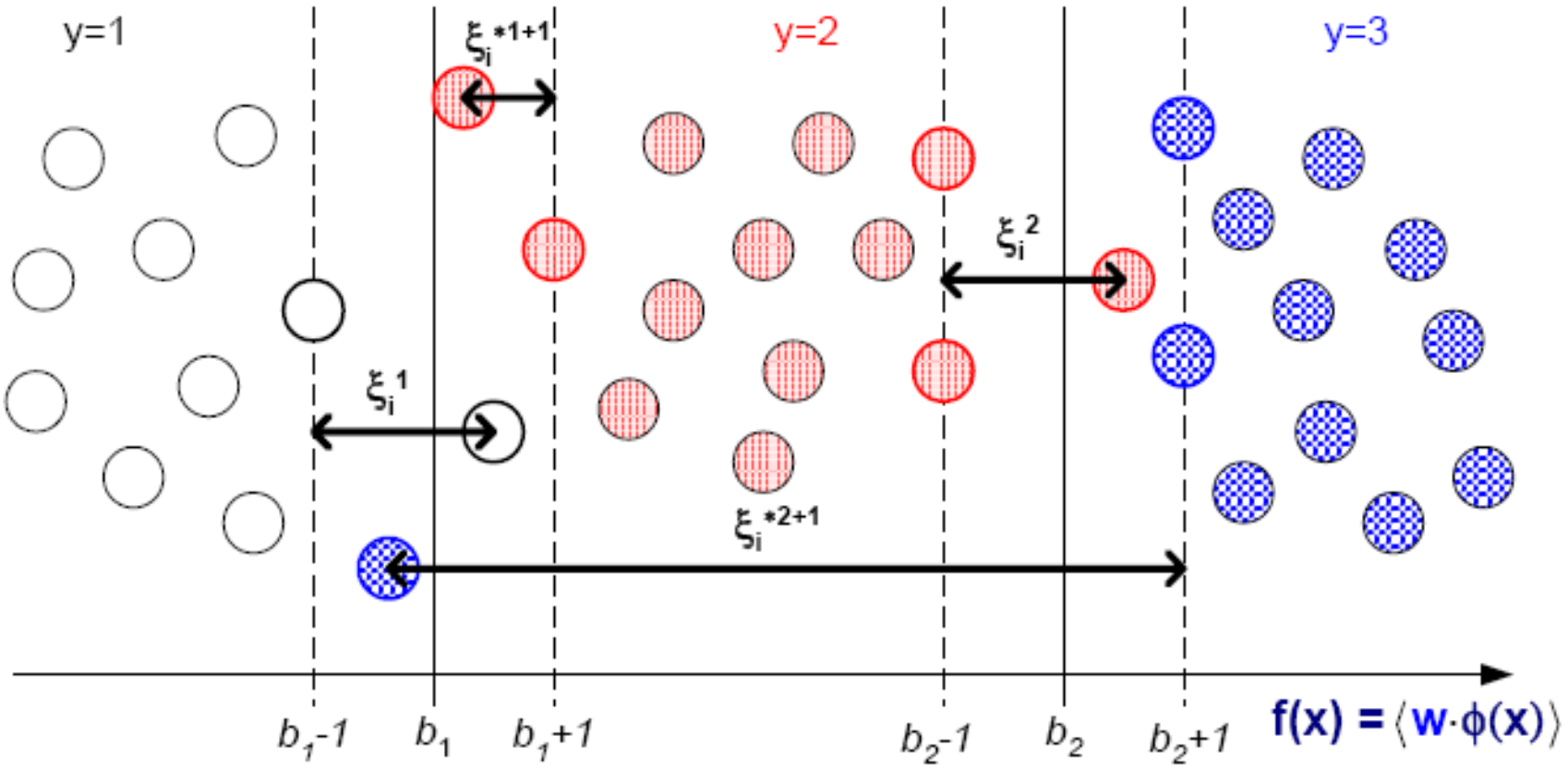
"Learning to rank"

- Assume a number of categories ***C*** of relevance exist
  - These are totally ordered: $c_1 < c_2 < \ldots < c_J$
  - This is the ordinal regression setup
- Assume training data is available consisting of document-query pairs represented as feature vectors ψ_i and relevance ranking c_i

- We could do *point-wise learning*, where we try to map items of a certain relevance rank to a subinterval
- But most work does *pair-wise learning*, where the input is a pair of results for a query, and the class is the relevance ordering relationship between them

Point-wise learning

- Goal is to learn a threshold to separate each



ΤΕΛΟΣ (μέρους) 19ου, 20ου Κεφαλαίου

Ερωτήσεις?

Χρησιμοποιήθηκε κάποιο υλικό από:

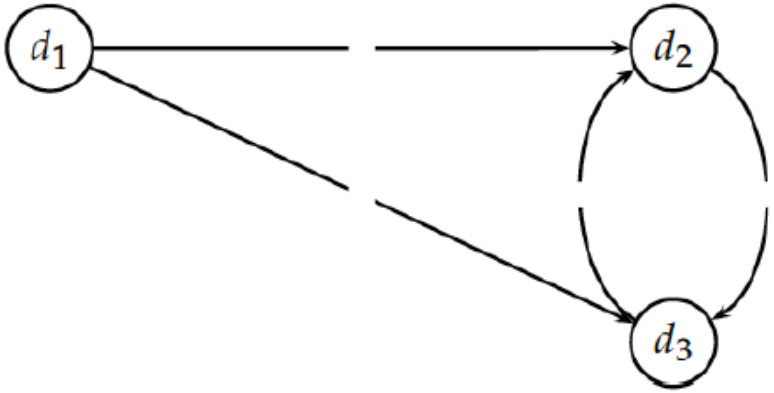
- ✓*Pandu Nayak and Prabhakar Raghavan, CS276:Information Retrieval and Web Search (Stanford)*
- ✓*Hinrich Schütze and Christina Lioma, Stuttgart IIR class*

Τι είναι η Ανάκτηση Πληροφορίας (Information Retrieval);

Τέλος Μαθήματος

ΑΣΚΗΣΕΙΣ

Άσκηση 21.22



PageRank

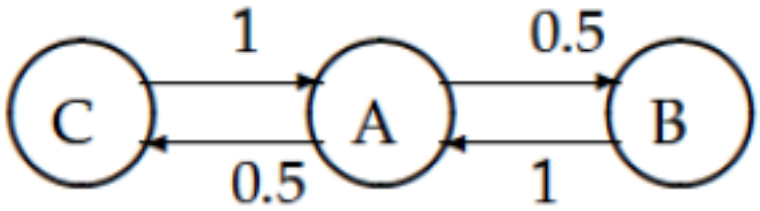
+teleporting με 0,8

HITS

Άσκηση 21.19

If all the hub and authority scores are initialized to 1, what is the hub/authority score of a node after one iteration?

Άσκηση 21.5

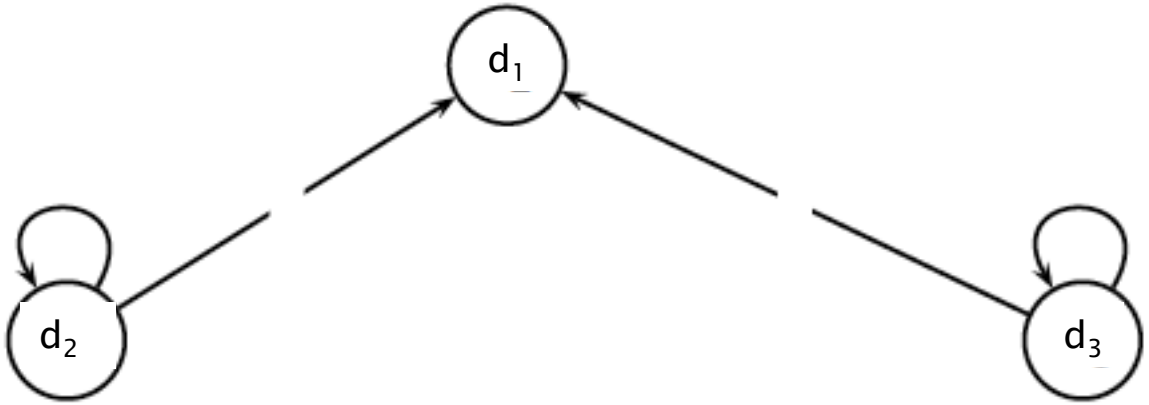


Πίνακα μετάβασης για Markov αλυσίδες
PageRank
Power iteration
Jumps με α = 0.8
Jumps με α = 0

Άσκηση 21.7

A user of a browser can, in addition to clicking a hyperlink on the page x she is currently browsing, use the back button to go back to the page from which she arrived at x. *Can such a use of back buttons be modelled as a Markov chain?* How would we model repeated invocations of the back button?

Άσκηση



PageRank
HITS

- **Exercise 21.11** Verify that the pagerank of the data in the following transition matrix (from book and lectures)

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0.02	0.02	0.88	0.02	0.02	0.02	0.02
d_1	0.02	0.45	0.45	0.02	0.02	0.02	0.02
d_2	0.31	0.02	0.31	0.31	0.02	0.02	0.02
d_3	0.02	0.02	0.02	0.45	0.45	0.02	0.02
d_4	0.02	0.02	0.02	0.02	0.02	0.02	0.88
d_5	0.02	0.02	0.02	0.02	0.02	0.45	0.45
d_6	0.02	0.02	0.02	0.31	0.31	0.02	0.31

is indeed

$\vec{x} = (0.05\, 0.04\, 0.11\, 0.25\, 0.21\, 0.04\, 0.31)$

ΤΕΛΟΣ Ασκήσεων

Ερωτήσεις?