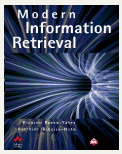
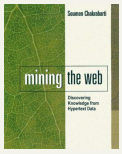
# Web Algorithmics
(The power of algorithms)

Paolo Ferragina
Dipartimento di Informatica
Università di Pisa

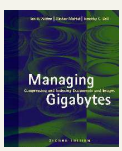
## Some references




Modern Information Retrieval
R. Baeza-Yates and B. Ribeiro-Neto, Addison-Wesley Publ., 1999.




Mining the Web: Discovering Knowledge from…
S. Chakrabarti, Morgan-Kaufmann Publishers, 2003.




Managing Gigabytes
A. Moffat, T. Bell e I. Witten, Kaufmann Publisher 1999.

Many slides on my web page

## Goal of a Search Engine

Retrieve docs that are "relevant" for the user query

- Doc: file word or pdf, web page, email, blog, e-book,...
- Query: paradigm "bag of words"

## Two main difficulties

**The Web:**

**Extracting "significant data" is difficult !!**

- Language and encodings: hundreds…
- Distributed authorship: SPAM, format-less,…
- Dynamic: in one year 35% survive, 20% untouched

**The User:**

**Matching "user needs" is difficult !!**

- Query composition: short (2.5 terms avg) and imprecise
- Query results: 85% users look at just one result-page
- Several needs: Informational, Navigational, Transactional

The "relevant" issue is subjective and time-varying

## Evolution of Search Engines

- First generation -- use only on-page, web-text data
  - Word frequency and language

1995-1997
AltaVista, Excite, Lycos, etc

- Second generation -- use off-page, web-graph data
  - Link (or connectivity) analysis
  - Anchor-text (How people refer to a page)

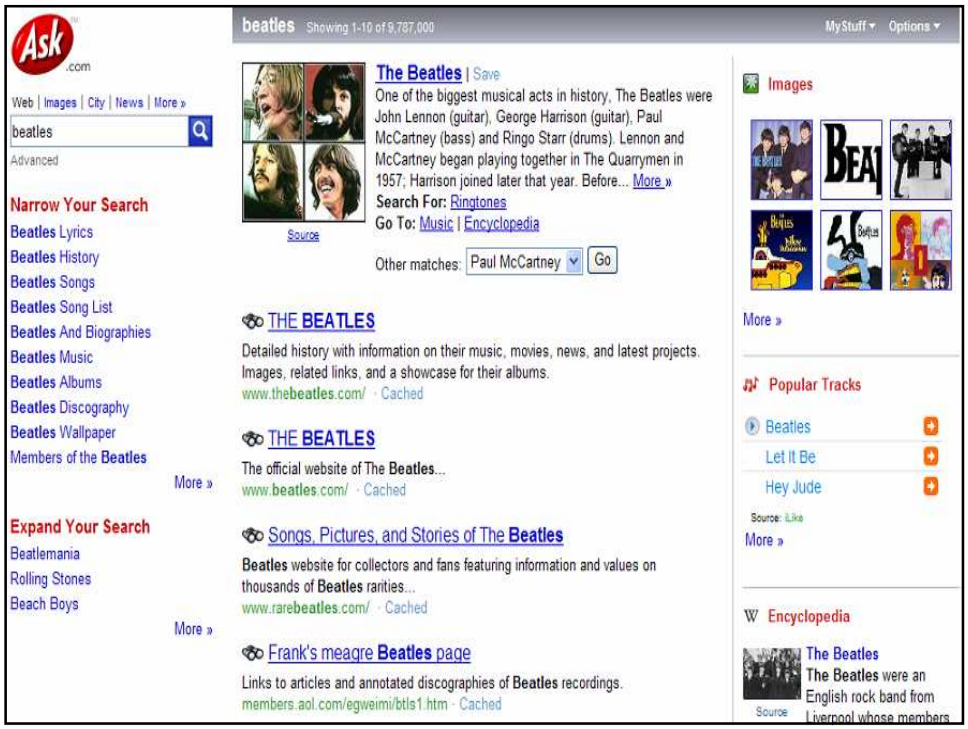
1998: Google

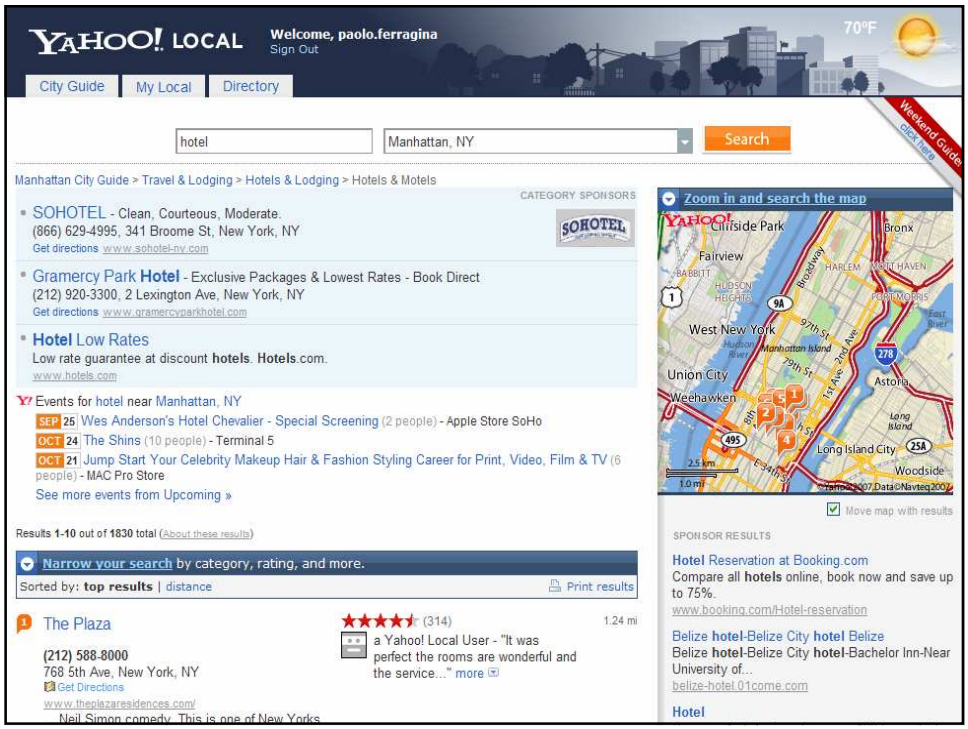
- Third generation -- answer "the need behind the query"
  - Focus on "user need", rather than on query
  - Integrate multiple data-sources
  - Click-through data

Google, Yahoo, MSN, ASK,……

**Fourth generation → Information Supply**
[Andrei Broder, VP emerging search tech, Yahoo! Research]


YAHOO!
Search:
Web Search
My Yahoo!
My Mail
Answers
Autos
Finance
Games
Groups
HotJobs
Maps
Mobile Web
Movies
Music
Personals
Real Estate
Shopping
Sports
Tech
Travel
TV
Yellow Pages
More Yahoo! Services
Small Business
Featured Services
Featured
Entertainment
Sports
Video
400 richest Americans
Hi, Paolo
Sign Out
Mail
Messenger
Radio
Weather
Local
Horoscopes
Yahoo! Travel
Plan Your Getaway
Today's Top Deals
Vacation Packages*
Huge Cruise Savings*
Hotel Deals Worldwide
*Info on taxes and fees
Search Travel:
Be a Better Movie Date
Watch movie trailers and clips
In the News
World
Local
Finance
Marketplace
Pulse - What Yahoos Are Into
All Stars: Most Viewed Cars


Ask.com
beatles
Showing 1-10 of 9,787,000
MyStuff
Options
Narrow Your Search
Beatles Lyrics
Beatles History
Beatles Songs
Beatles Song List
Beatles And Biographies
Beatles Music
Beatles Albums
Beatles Discography
Beatles Wallpaper
Members of the Beatles
Expand Your Search
Beatlemania
Rolling Stones
Beach Boys
The Beatles | Save
Search For: Ringtones
Go To: Music | Encyclopedia
Other matches: Paul McCartney
THE BEATLES
THE BEATLES
Songs, Pictures, and Stories of The Beatles
Frank's meagre Beatles page
Images
Popular Tracks
Beatles
Let It Be
Hey Jude
Encyclopedia
The Beatles

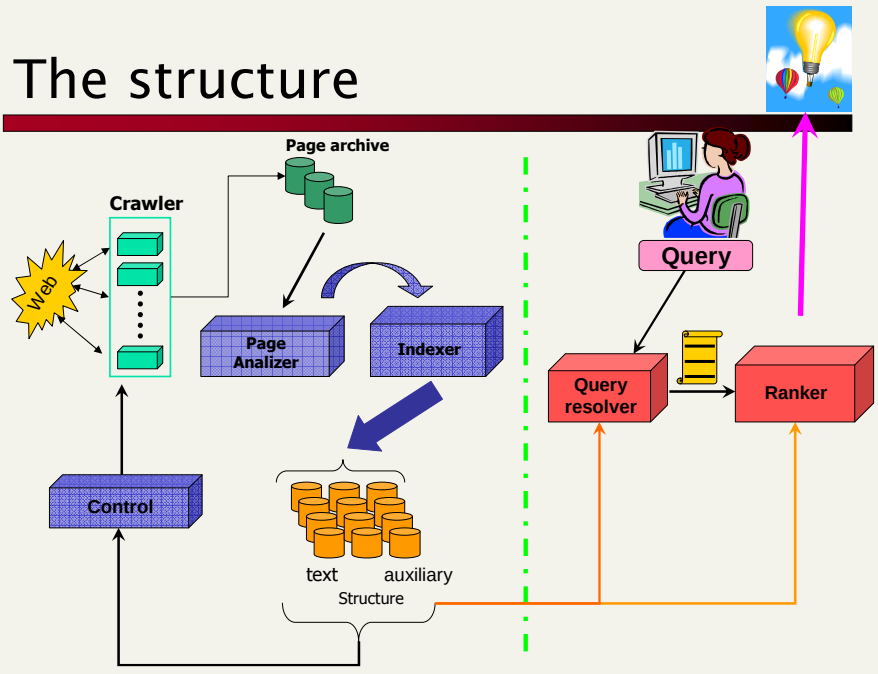


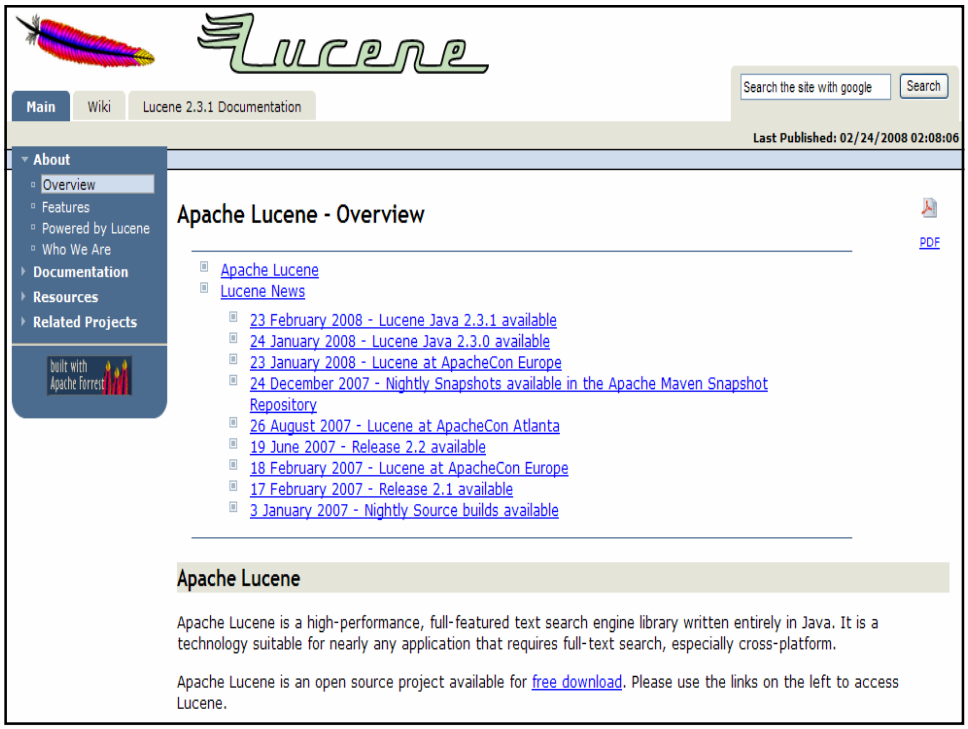

# This is a search engine!!!

# Web Algorithmics
(The power of algorithms)

The structure of a Search Engine

## The structure






# Web Algorithmics
## (The power of algorithms)

Algorithmic Issues:
I/O and space

## Algorithms vs. Hardware features

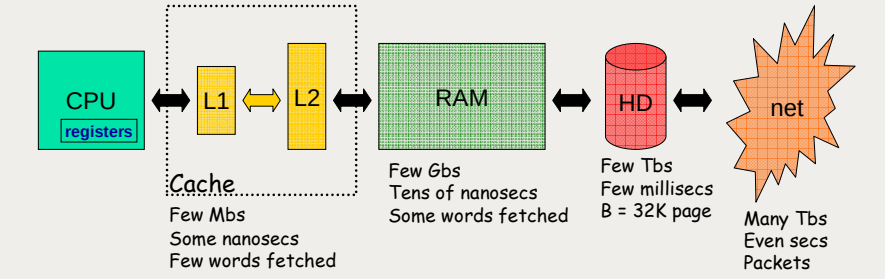
- We have three types of algorithms:
  - $T_1(n) = n$, $T_2(n) = n^2$, $T_3(n) = 2^n$

... and assume that 1 step = 1 time unit

- How many input data n, each algorithm may process within t time units?
  - $n_1 = t$,
- What about a k-times faster processor?
  ...or, what is n, when the time units are k*t ?
  - $n_1 = k * t$,

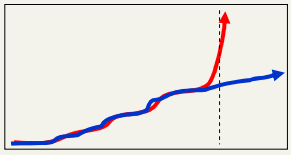
## Algorithm Inadequacy

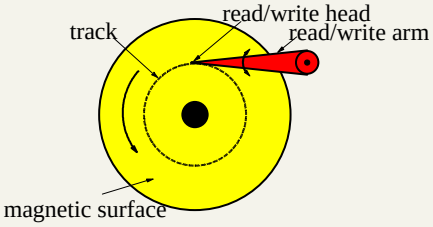
- Importance of scalability/efficiency: Algorithmics is at the core
- Traditional algorithmics uses a simple machine model




**You should be space/IO-aware "programmers"**

# I/O-conscious Algorithms






*"The difference in speed between modern CPU and disk technologies is analogous to the difference in speed in sharpening a pencil using a sharpener on one's desk or by taking an airplane to the other side of the world and using a sharpener on someone else's desk."* (D. Comer)

**Spatial locality vs Temporal locality**

# Space-conscious Algorithms

IBM Research

## Conclusions

Systems should automatically compress data whenever the benefits of storing or transmitting the compressed data outweigh the costs

- It's time to "teach" systems how to do this
  - Don't stick just to "vanilla" compression

15 Toward Ubiquitous Compression © 2004 IBM Corporation

## Problem: Indexing

- Consider Wikipedia En:
  - Collection size ≈ *10* Gbytes
  - # docs ≈ *4 * 10^6*
  - #terms in total > *1 billion* (avg term len = 6 chars)
  - #terms distinct = *several millions*
- Which data structure do we build in order to support word-based searches ?

## DB-based solution: Term-Doc matrix

#docs ≈ 4M

#terms > 1M

| | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth |
|---|---|---|---|---|---|---|
| Antony | 1 | 1 | 0 | 0 | 0 | 1 |
| Brutus | 1 | 1 | 0 | 1 | 0 | 0 |
| Caesar | 1 | 1 | 0 | 1 | 1 | 1 |
| Calpurnia | 0 | 1 | 0 | 0 | 0 | 0 |
| Cleopatra | 1 | 0 | 0 | 0 | 0 | 0 |
| mercy | 1 | 0 | 1 | 1 | 1 | 1 |
| worser | 1 | 0 | 1 | 1 | 1 | 0 |

1 if play contains word, 0 otherwise

Space ≈ 4Tb !

# Compression vs Speed

- **M** = memory size, **N** = problem size
- **T(n)** = time complexity of an algorithm using linear space
- **p** = fraction of memory accesses [0,3÷0,4 (Hennessy-Patterson)]
- **C** = cost of an I/O [$10^5 \div 10^6$ (Hennessy-Patterson)]

If N=(1+f)M, then the Δ–avg cost per step is:

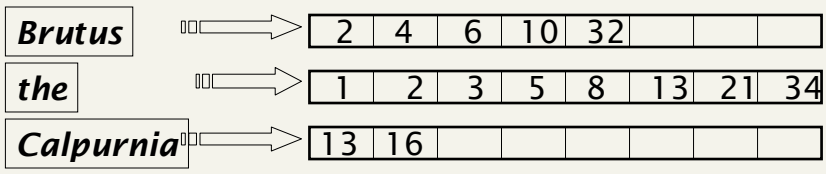
$$C * p * f/(1+f)$$

**This is at least** $10^5 * f/(1+f)$

**If we fetch B ≈ 4Kb in time C, and algo uses all of them:**

$$(1/B) * (p * f/(1+f) * C) \approx 10^2 * f/(1+f)$$

# Current solution: Inverted index




- A term like ***Calpurnia*** may use $\log_2$ N bits per occurrence
- A term like ***the*** should take about 1 bit per occurrence

Currently they get 30÷50% original text

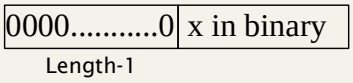
## Gap-coding for postings

- Sort the docIDs
- Store gaps between consecutive docIDs:
  - ***Brutus***: 33, 47, 154, 159, 202 …
    33, 14, 107, 5, 43 …

Two advantages:

- Space: store smaller integers (clustering?)
- Speed: query requires just a scan

## γ–code for integer encoding

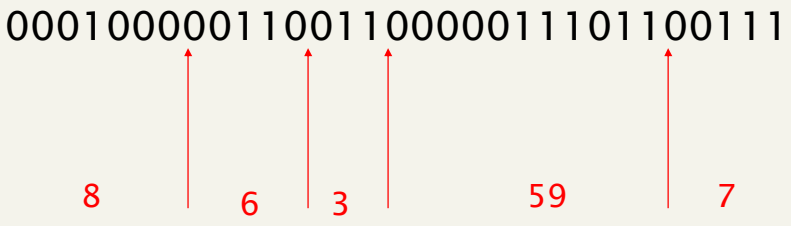



- $x > 0$ and Length = $\lfloor \log_2 x \rfloor + 1$

e.g., 9 represented as <000,1001>.

- γ–code for $x$ takes $2 \lfloor \log_2 x \rfloor + 1$ bits
  (ie. factor of 2 from optimal)

## It is a prefix-free encoding…




## Variable-byte codes

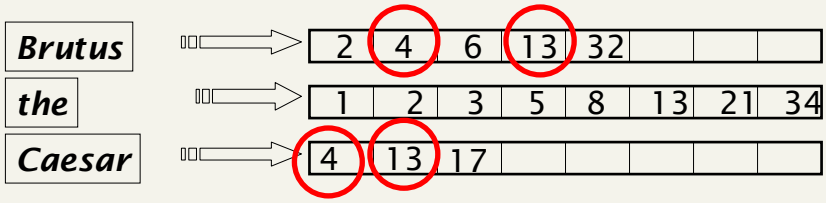
- Wish to get very fast (de)compress → byte-align

e.g., $v=2^{14}+1$ → *binary(v)* = 100000000000001

1*0000000*1 10000000 00000001

Note: We waste 1 bit per byte, and avg 4 for the first byte.

## Query processing

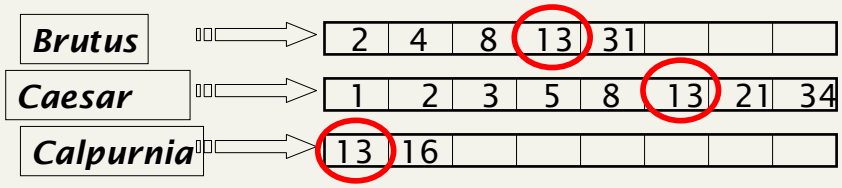



1) Retrieve all pages matching the query

## Some optimization

Best order for query processing ?

- Shorter lists first…

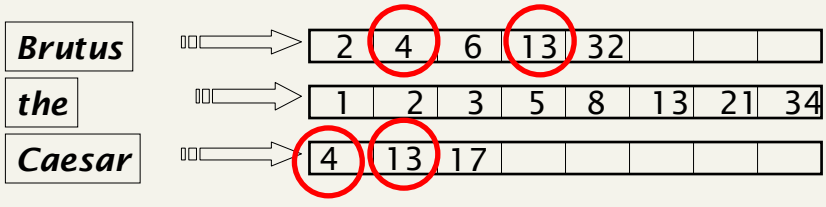



Query: ***Brutus*** *AND* ***Calpurnia*** *AND* ***Caesar***

# Phrase queries

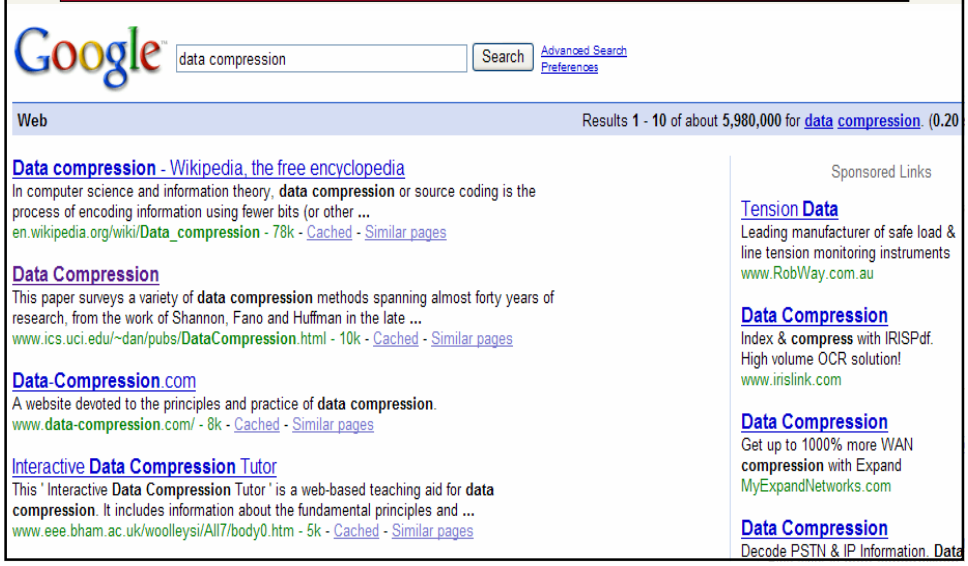
- Expand the posting lists with word positions
  - ***to***:
    - *2*:1,17,74,222,551; *4*:8,16,190,429,433; *7*:13,23,191; ...
  - ***be***:
    - *1*:17,19; *4*:17,191,291,430,434; *5*:14,19,101; ...
- Larger space occupancy, about 4 times more

# Query processing




1) Retrieve all pages matching the query
2) Order pages according to various scores:
   - Term position & freq (body, title, anchor,...)
   - Link popularity
   - User clicks or preferences

# Generating the snippets !




# An interesting issue

Compressed DB of Web pa[illegible] → Snippets !!

- Gbs/Tbs of data
- Data are heterog[illegible]
- Support (random[illegible]

Two main ([illegible]

- Gzip e[illegible]
- Huffword[illegible]

Whatever is your idea, it must be SCALABLE

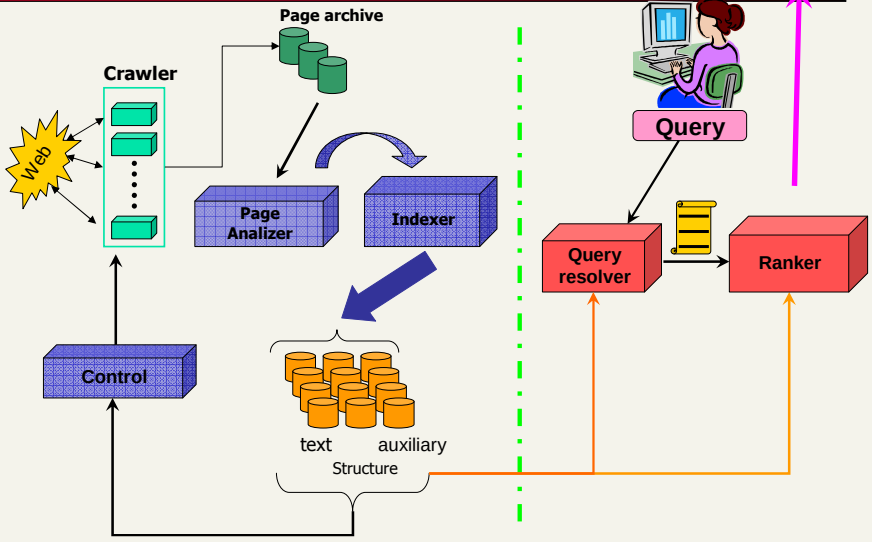
Various issues [illegible]

1. Clustering
2. Parsing = Dictionary [illegible]
3. (Optimal) Model in lim[illegible]d space

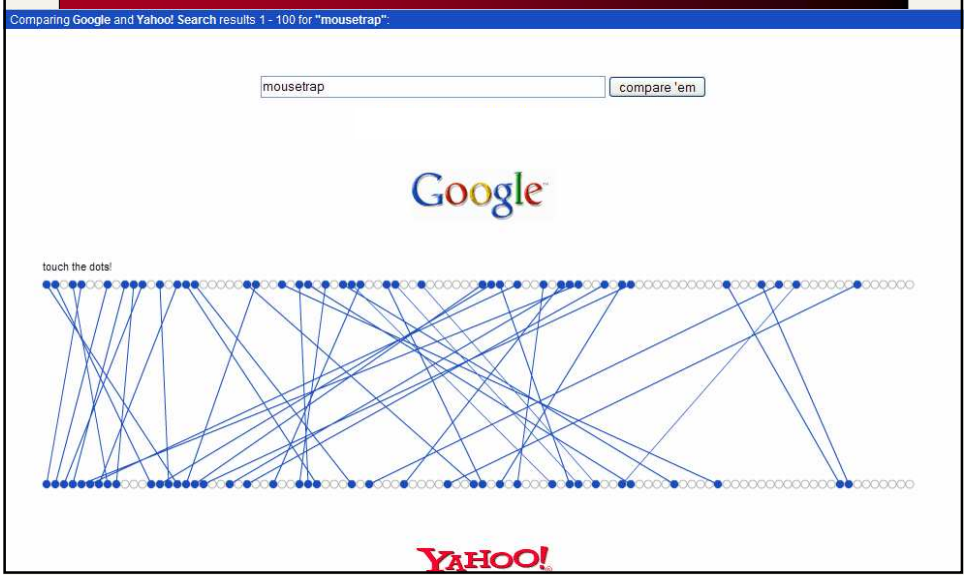
On 47Gb of UK-pages [UniMi]

- ✓ Huffword ≈ 21Gb
- ✓ Gzip ≈ 12Gb
- ✓ Bzip ≈ 8Gb

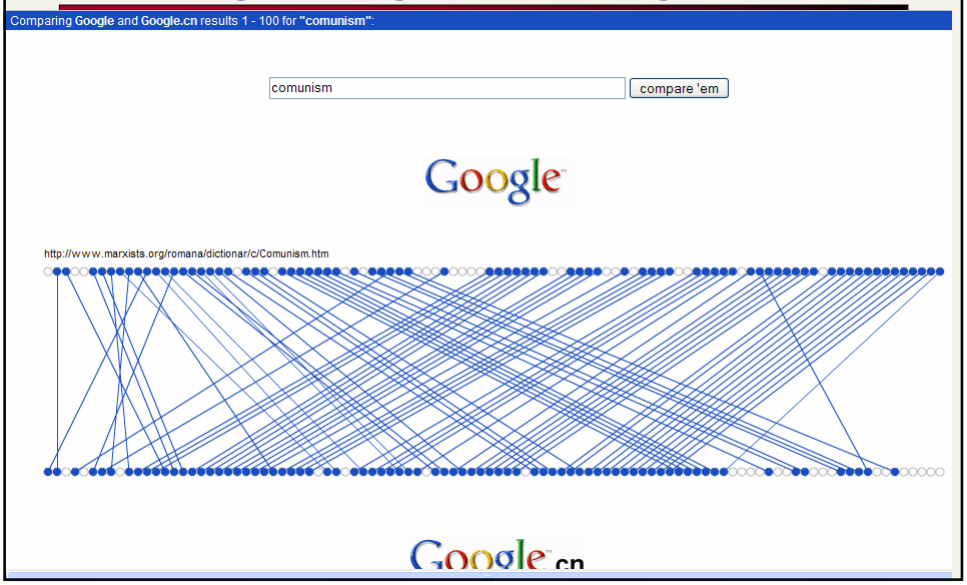
# The structure




# The big fight: find the best *ranking*...



## Ranking: Google *vs* Google.cn




# Web Algorithmics
(The power of algorithms)

## Text-based Ranking
(1° generation)

# A famous "weight": tf-idf

$$w_{t,d} = tf_{t,d} \times \log(n / n_t)$$

$tf_{t,d}$ = Frequency of term $t$ in doc $d$ = $\#occ_t / |d|$

$$idf_t = \log\left(\frac{n}{n_t}\right)$$

where $n_t$ = #docs containing term $t$

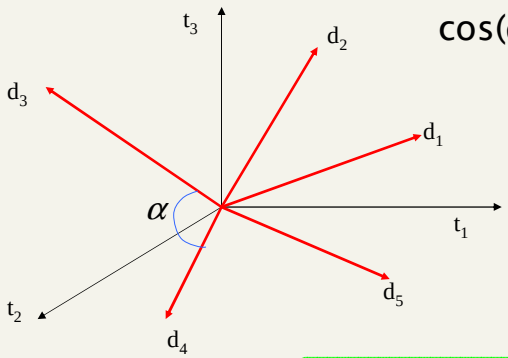
n = #docs in the indexed collection

| | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth |
|---|---|---|---|---|---|---|
| Antony | 13,1 | 11,4 | 0,0 | 0,0 | 0,0 | 0,0 |
| Brutus | 3,0 | 8,3 | 0,0 | 1,0 | 0,0 | 0,0 |
| Caesar | 2,3 | 2,3 | 0,0 | 0,5 | 0,3 | 0,3 |
| Calpurnia | 0,0 | 11,2 | | | | |
| Cleopatra | 17,7 | 0,0 | | | | |
| mercy | 0,5 | 0,0 | | | | |
| worser | 1,2 | 0,0 | 0,6 | 0,6 | 0,6 | 0,0 |

Vector Space model

# A graphical example

**Easy to Spam**




$\cos(\alpha) = v \bullet w / ||v|| * ||w||$

Sophisticated algos to find top-k docs for a query Q

The user query is a very short doc

*Postulate*: Documents that are "close together" in the vector space talk about the same things. Euclidean distance sensible to vector length !!

## Approximate top-k results

- Preprocess: Assign to each term, its *m* best documents

| | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth |
|---|---|---|---|---|---|---|
| Antony | 13.1 | 11.4 | 0.0 | 0.0 | 0.0 | 0.0 |
| Brutus | 3.0 | 8.3 | 0.0 | 1.0 | 0.0 | 0.0 |
| Caesar | 2.3 | 2.3 | 0.0 | 0.5 | 0.3 | 0.3 |
| Calpurnia | 0.0 | 11.2 | 0.0 | 0.0 | 0.0 | 0.0 |
| Cleopatra | 17.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| mercy | 0.5 | 0.0 | 0.7 | 0.9 | 0.9 | 0.3 |
| worser | 1.2 | 0.0 | 0.6 | 0.6 | 0.6 | 0.0 |

- Search:
  - If |Q| = *q* terms, merge their preferred lists ($\leq mq$ answers).
  - Compute COS between Q and these docs, and choose the top *k*.

Need to pick $m>k$ to work well empirically.

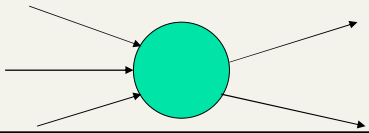
Now SE use tf-idf PLUS PageRank (PLUS other weights)

## Web Algorithmics
(The power of Algorithms)

## Link-based Ranking
(2° generation)

## Query-independent ordering

- First generation: using link counts as simple measures of popularity.
  - Undirected popularity:
    - Each page gets a score given by the number of in-links plus the number of out-links (es. 3+2=5).
  - Directed popularity:
    - Score of a page = number of its in-links (es. 3).

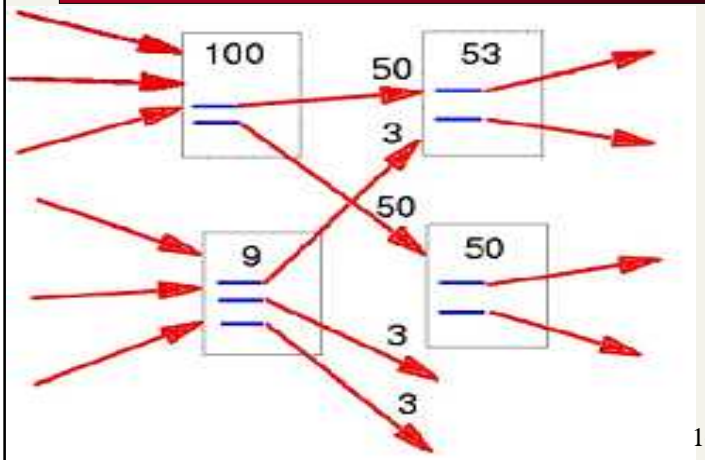


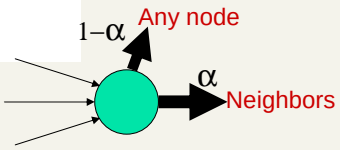
Easy to SPAM

## Second generation: PageRank

- Each link has its own importance!!
- PageRank is
  - independent of the query
  - many interpretations…

## Basic Intuition…




## Google's Pagerank

$$r(i) = \alpha \cdot \sum_{j \in B(i)} \frac{r(j)}{\#out(j)} + (1-\alpha) \cdot \frac{1}{N}$$

**Fixed value**

**Principal eigenvector**

$$r = \left[\alpha \Pi^T + (1-\alpha) e e^T\right] \times r$$

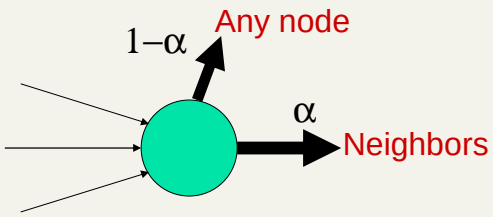
**B(i)** : set of pages linking to i.
**#out(j)** : number of outgoing links from j.

$$\Pi_{i,j} = \begin{cases} \frac{1}{\#out(i)} & i \to j \\ 0 & else \end{cases}$$

## Three different interpretations

- Graph (intuitive interpretation)
  - Co-citation
- Matrix (easy for computation)
  - Eigenvector computation or a linear system solution
- Markov Chain (useful to prove convergence)
  - a sort of Usage Simulation

"*In the steady state*" each page has a long-term visit rate
- use this as the page's score.




## Pagerank: use in Search Engines

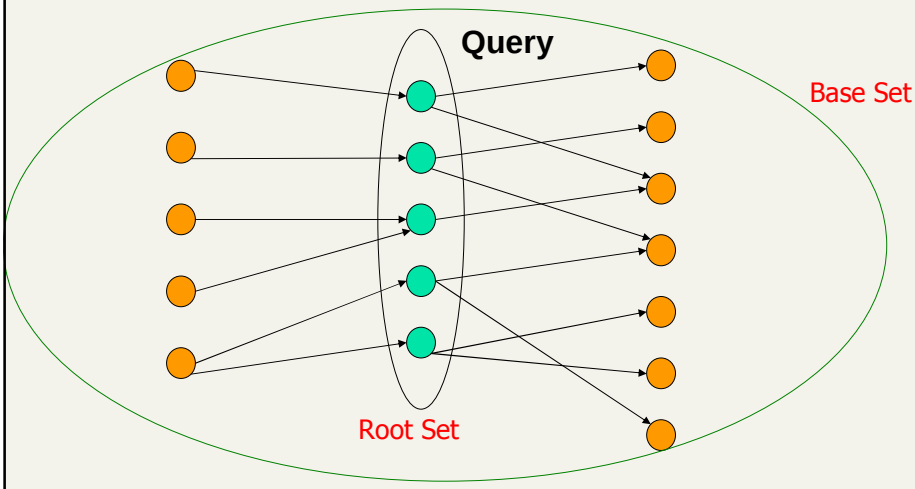
- Preprocessing:
  - Given graph of links, build matrix **P**
  - Compute its **principal eigenvector r**
  - r[i] is the pagerank of page *i*

We are interested in the *relative order*

- Query processing:
  - Retrieve pages containing query terms
  - Rank them by their Pagerank

The final order is query-*independent*

## HITS: Hypertext Induced Topic Search

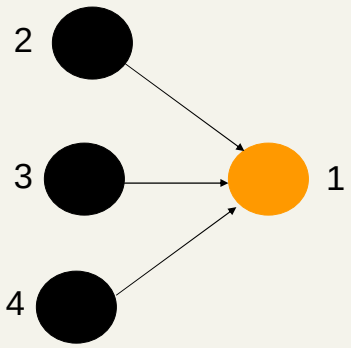



## Calculating HITS

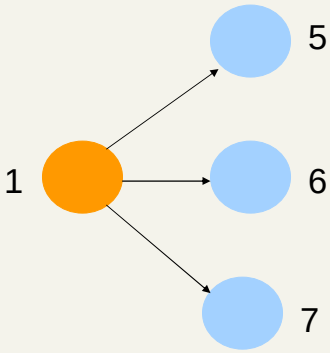
- It is query-*dependent*

- Produces two scores per page:
  - **Authority score**: a *good authority* page for a topic is *pointed* to by many good hubs for that topic.
  - **Hub score:** A *good hub* page for a topic *points* to many authoritative pages for that topic.

## Authority and Hub scores




a(1) = h(2) + h(3) + h(4)




h(1) = a(5) + a(6) + a(7)

## HITS: Link Analysis Computation

$$\left.\begin{aligned} a &= A^T h \\ h &= Aa \end{aligned}\right\} \Rightarrow \begin{aligned} a &= A^T A a \\ h &= A A^T h \end{aligned}$$

Where

a: Vector of Authority's scores

h: Vector of Hub's scores.

A: Adjacency matrix in which $a_{i,j} = 1$ if i→j

Thus, **h** is an eigenvector of $\mathbf{AA^t}$

**a** is an eigenvector of $\mathbf{A^tA}$

## Weighting links

Weight more if the query occurs in the neighborhood of the link (e.g. anchor text).

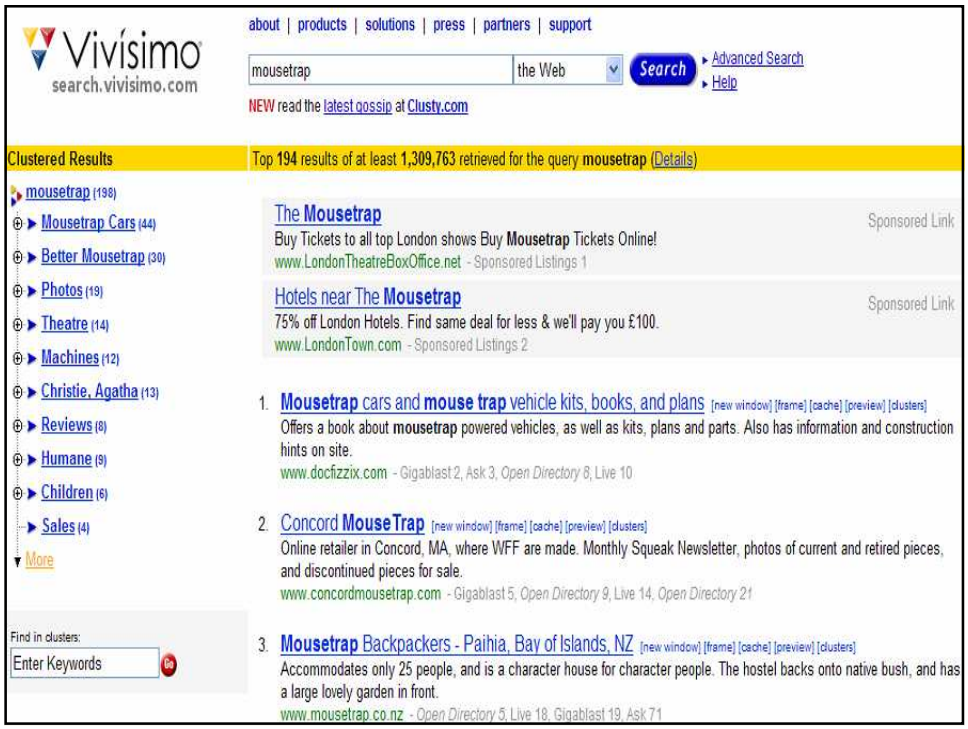
$$h(x) \leftarrow \sum_{x \mapsto y} a(y) \qquad a(x) \leftarrow \sum_{y \mapsto x} h(y)$$

$$h(x) = \sum_{x \mapsto y} w(x,y) \cdot a(y) \qquad a(x) = \sum_{y \mapsto x} w(x,y) \cdot h(y)$$

## Web Algorithmics
(The power of algorithms)

Clustering of Search Results

(a different approach…)




# Web-Snippet Hierarchical Clustering

- The folder hierarchy must be formed
  - "**on-the-fly from the snippets**": because it must adapt to query results without any costly remote access to the original web pages or documents
  - "**and his folders may overlap**": because a snippet may deal with multiple themes
- The folder labels must be formed
  - "**on-the-fly from the snippets**" because labels must capture the **potentially unbounded themes** of the results without any costly remote access to the original web pages or documents.
  - "**and be intelligible sentences**" because they must facilitate the post-navigation

Much research
and several softwares


# Two examples

**Clusters**

>>Personalized >>Unpersonalized >>Uncheck All
>>Expand All >>Collapse All >>Show Number Docs

- Soap
  - Protocol
  - Make
  - Natural Soap
  - Oils
  - Operas
  - Handcrafted Soap
  - Services Soap
    - Services For Soap
    - Interoperability Lab
    - Sqldata Soap
    - Directory For Soap
  - Soap Messaging
  - Php Soap
  - Liquid Soap
    - Personal Care Products
    - Liquid Soap
  - Detergent Association
  - Userland

more...

**Clusters**

>>Personalized >>Unpersonalized >>Uncheck All
>>Expand All >>Collapse All >>Show Number Docs

- Apache
  - Server
  - Project
  - Site
  - Software
  - Indians Histories
    - Apache Tribe
    - Apache Indians
  - Apache Webserver
    - Interface To Openssl
    - Apache Webserver
  - Point Observatory
  - Apache Python

more...

>>Personalized >>Unpersonalized >>Uncheck All
>>Expand All >>Collapse All >>Show Number Docs

## SnakeT's main features [Ferragina-Gullì, WWW 2005]

☺ Labels are *gapped* sentences of variable length

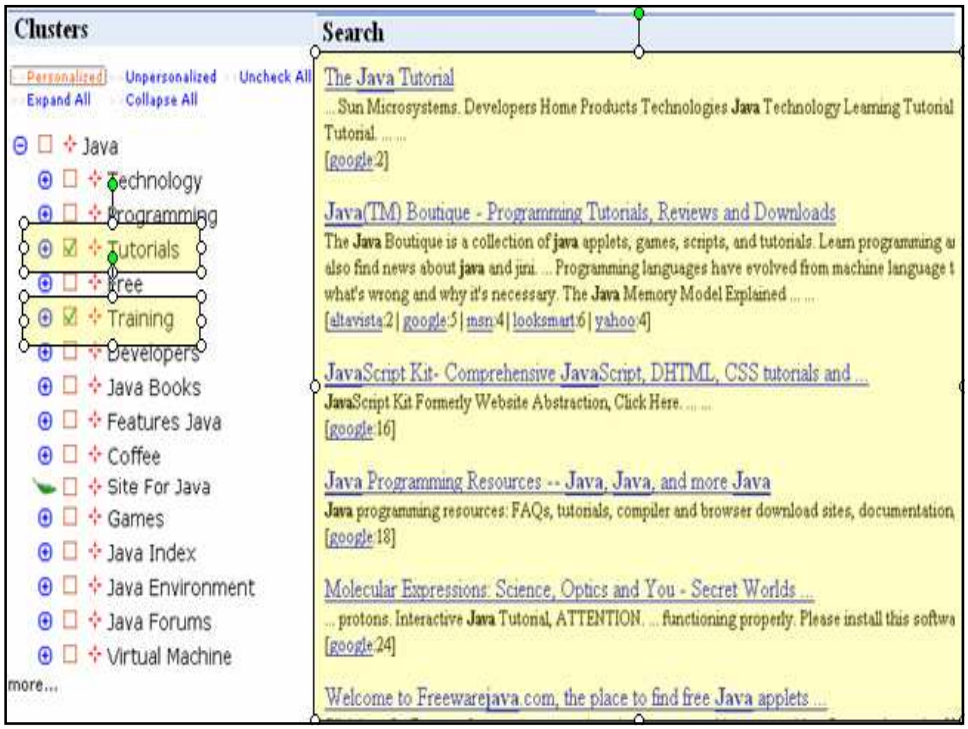
- To match sentences which are "almost the same"

☺ 2 knowledge bases for ranking/choosing the labels

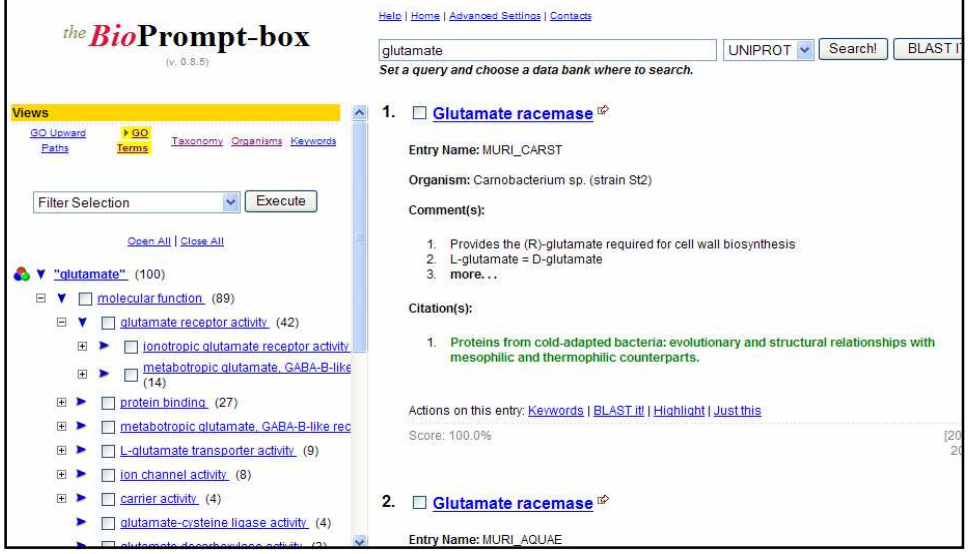
- DMOZ + Text Anchors

☺ Hierarchy formation uses folder labels and coverage

- Greedy approach to Graph Bipartite-&-Weighted Matching Problem



# BioPromptBox [Ferragina *et al*, *BioInformatics 2007*]

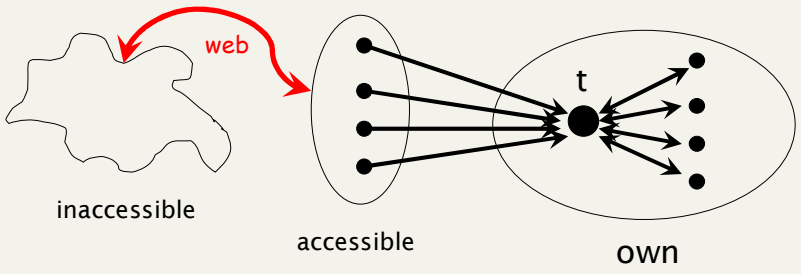



# An interesting issue

# Web Algorithmics
(The power of algorithms)

Web SPAM
(the real competitor…)

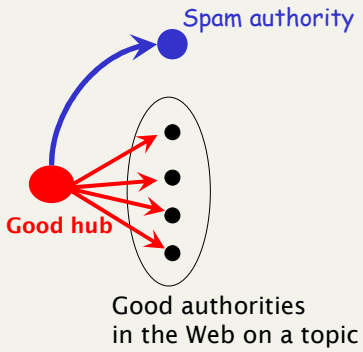
## Spamming PageRank




It is NOT easy to spam

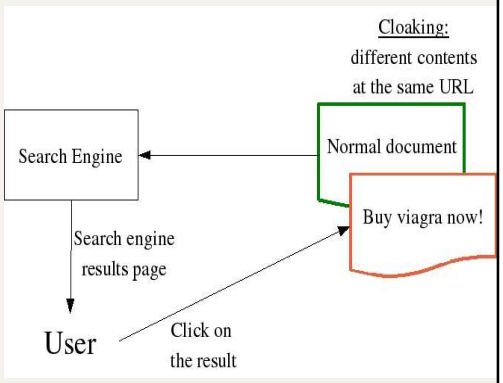
## Spamming HITS

It is easy to create a "fake" authority




## Adversarial IR

- ☑ Link spam
- ☑ Content spam
- ☑ Cloacking
- ☑ Click fraud
- ☑... many others...



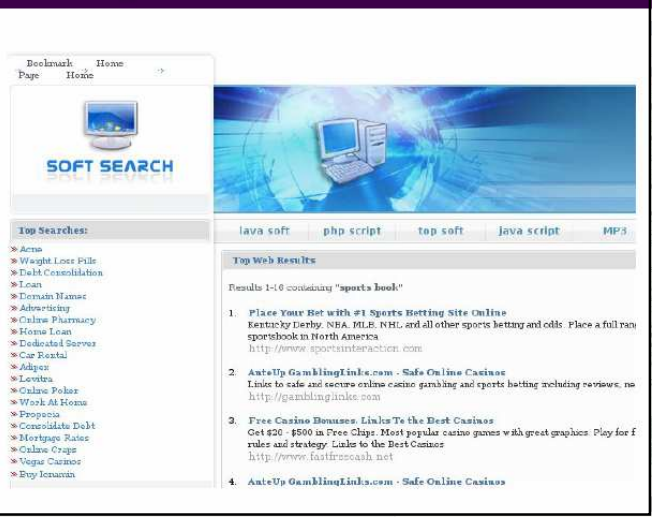
## SPAM costs

- ☑ For the user
  - ❖ Lower precision for some queries
- ☑ For the search engine
  - ❖ Waste storage space, network resources, processing cycle costs, authoritativeness,
- ☑ For the publisher
  - ❖ Resources invested in cheating and not in producing innovative/qualitative content

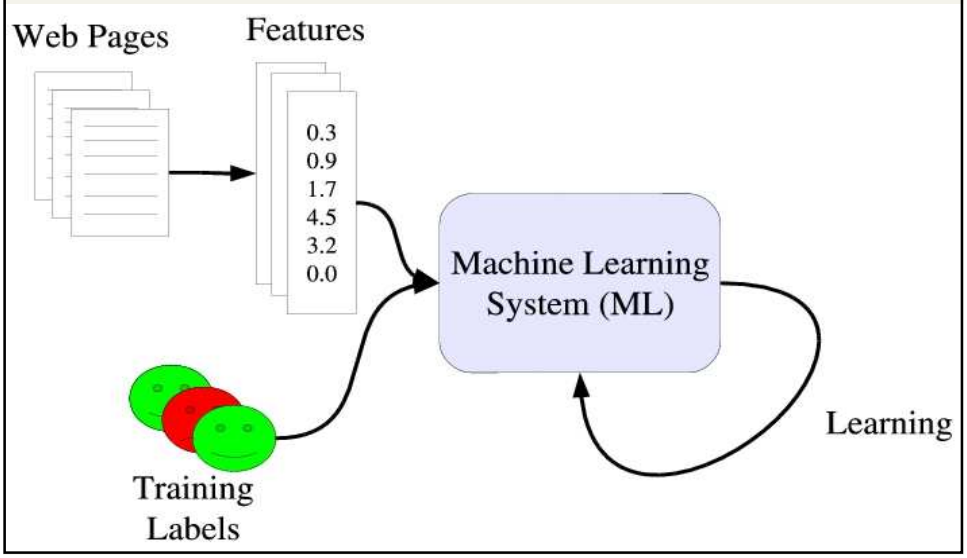
## Some examples of SPAM…

Made for advertising


Search engine?



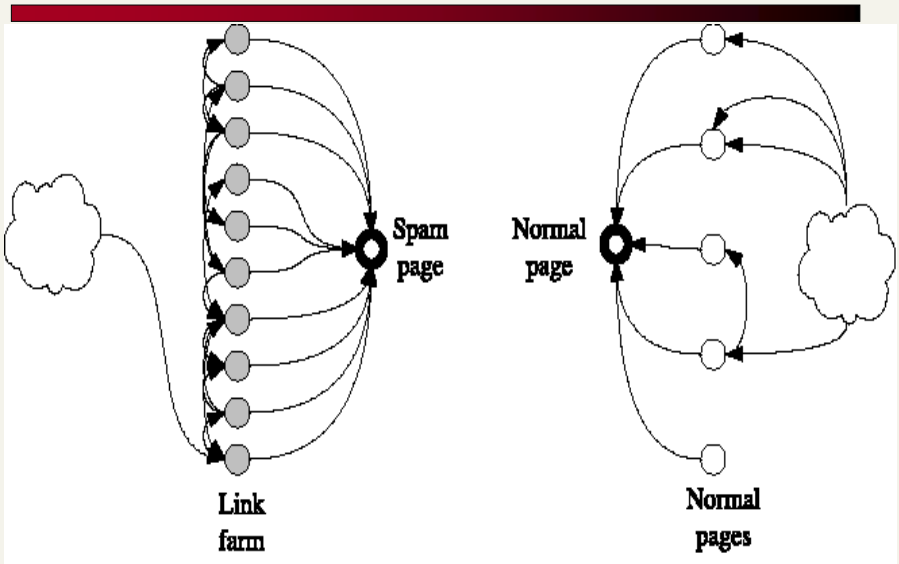
## Typical approach




## Content-based features

- ☑ number of out/in links
- ☑ number of words in the body/title
- ☑ avg word length
- ☑ fraction of anchor text or visible text
- ☑ compression rate

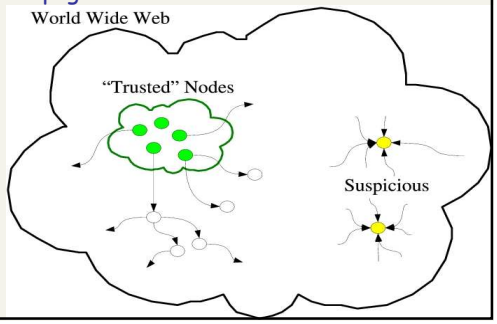
## Link-based features




## Recent approaches

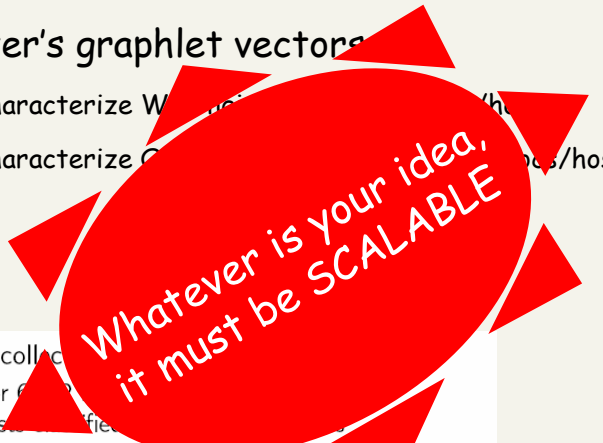
- ☑ Query-Log graph (spammer goals vs. their techniques)
  - ☑ Syntactic: Query-dictionary size and popularity
  - ☑ Semantic: Topic distribution of the query neighbors

Propagation



## An interesting issue

- ☑ Pinter's graphlet vectors
  - ☑ Characterize W
  - ☑ Characterize C /hosts




- Public spam coll
  - Labels for
  - 2,725 hos
  - 3,106 automatically co mal (.ac.uk, .sch.uk, .gov.uk, .m .nhs.uk or .police.uk)
  - **http://www.yr-bcn.es/webspam/**

# Web Algorithmics
(The power of Algorithms)

## Web Advertising
(the real business…)

## Why search is free ?

At least 85% users arrive to a site from a SE

1/3 users believe that "the results of a query are the best place where to buy things" !!

"There is a new type of economics that has emerged and that the world doesn't understand"

"Web usage data is an amazing leading indicator because it tells you where intent is heading"

U. Fayyad, Yahoo Chief Data Officer

## Search Engines *vs* Advertisement

- First generation -- use only on-page, web-text data
  - Word frequency and language

Pure search *vs* Paid search

- Second generation -- use off-page, web-graph data
  - Link (or connectivity) analysis
  - Anchor-text (How people refer to a page)

Ads show on search (who pays more), Goto/Overture

- Third generation -- answer "the need behind the query"
  - Focus on "user need", rather than on query
  - Integrate multiple data-sources
  - Click-through data

2003 Google/Yahoo New model

All players now have:
SE, Adv platform + network

# Typical approach…

Socio-demo | Geographic | Contextual




PROPERTY CENTRIC


YAHOO! NOTIZIE
hotels.com
YAHOO! ANSWERS
chiedi.
rispondi.
scopri.
YAHOO! FINANZA
YAHOO! Giochi
YAHOO!


**Relevant** messages
to users whose behavior
shows a clear interest
for a product or a service
YAHOO!

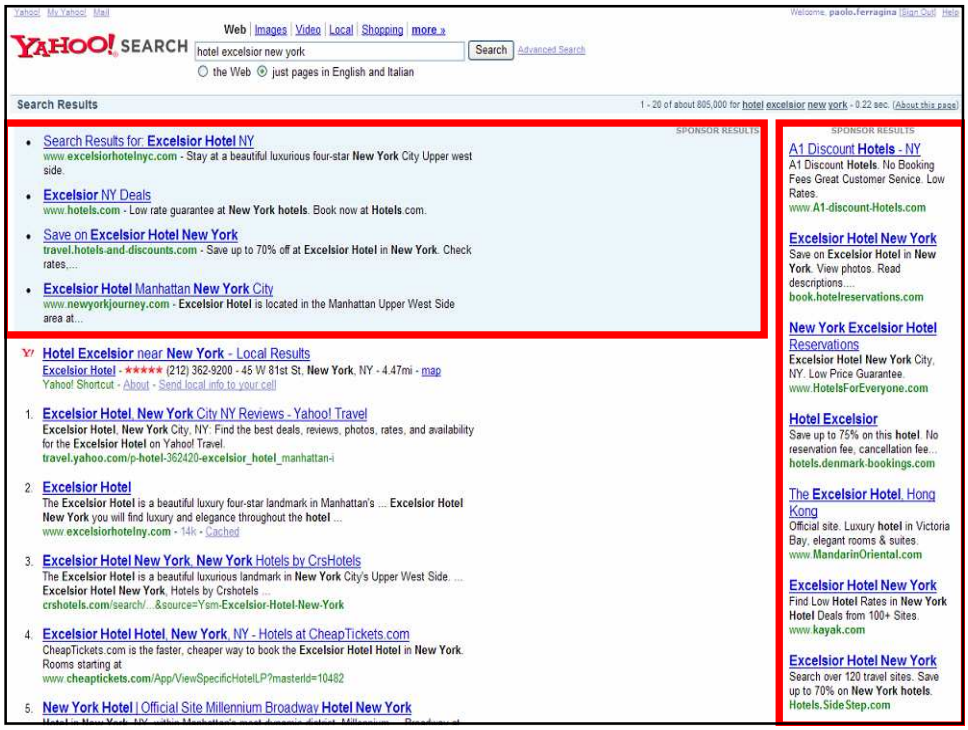
# Two approaches

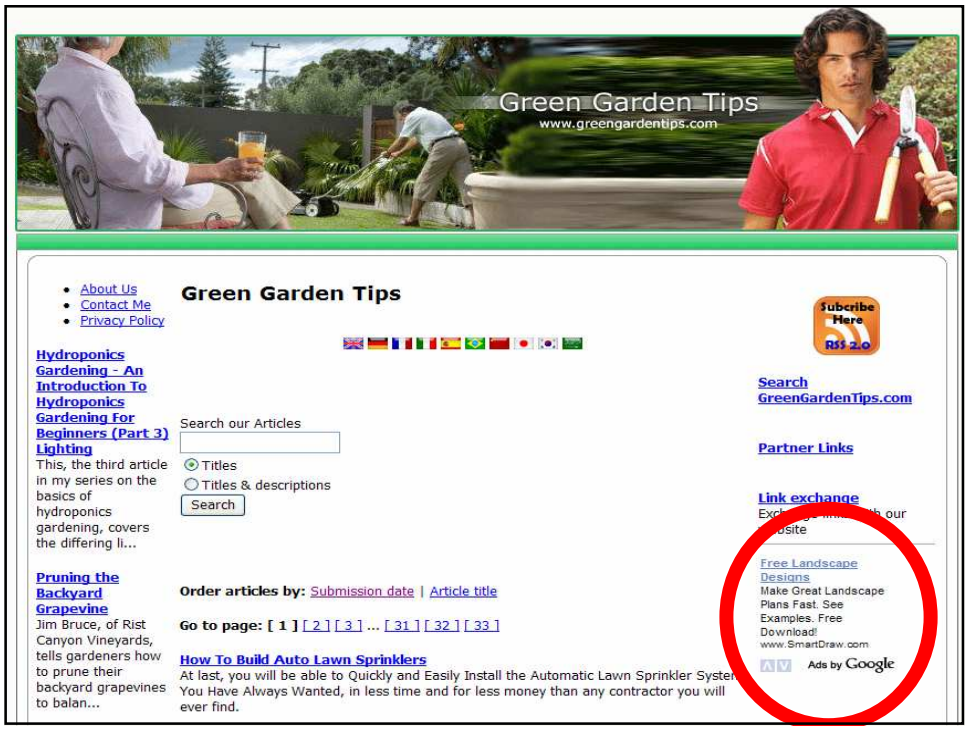
- Sponsored search: Ads driven by search keywords

AdWords

- Context match: Ads driven by the content of a web page

AdSense



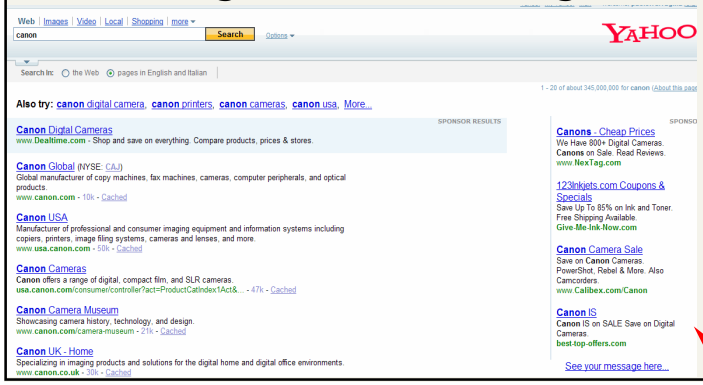



# How does it work ?

1) Match Ads to query or context
2) Order the Ads
3) Pricing on a click-through

IR

Econ



## How do we order Ads ?

- Bid ordering [initially Goto, Overture,...]
- Revenue ordering [now all]
  - by decreasing $R_x = bid_x * CTR_x$

Does Revenue Ordering maximize revenues?

NO – People react to the ordering, by changing their bid bahavior !!

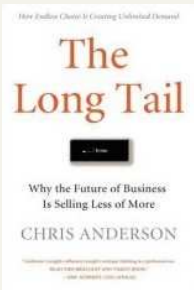
Competitive Market: many advertisers, users and the central service [Mechanism Design]

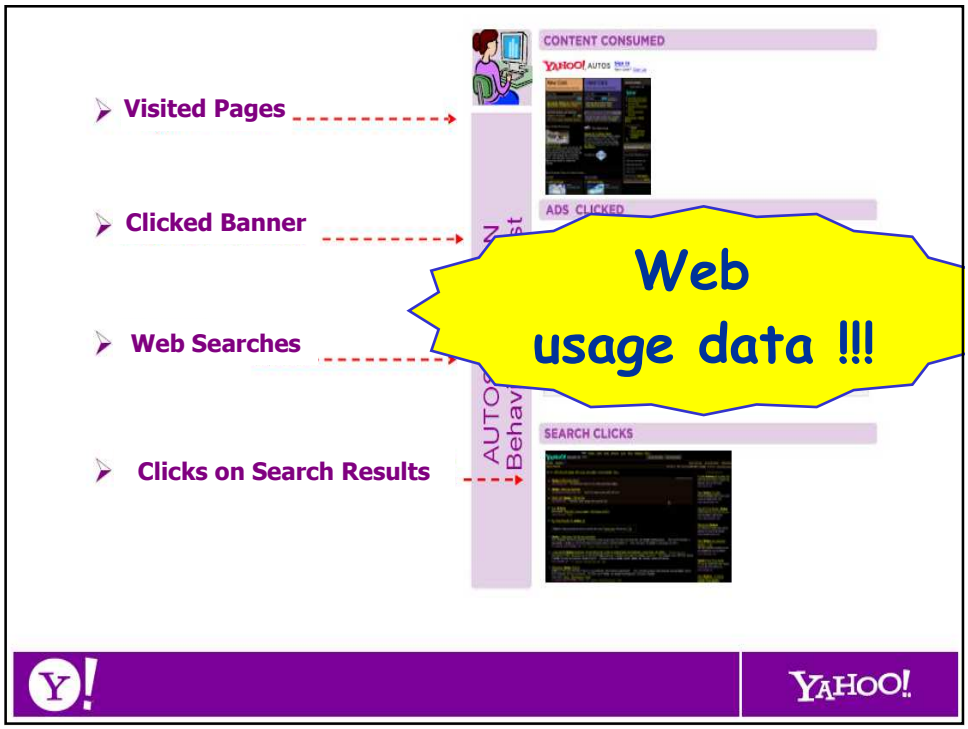
## The new scenario

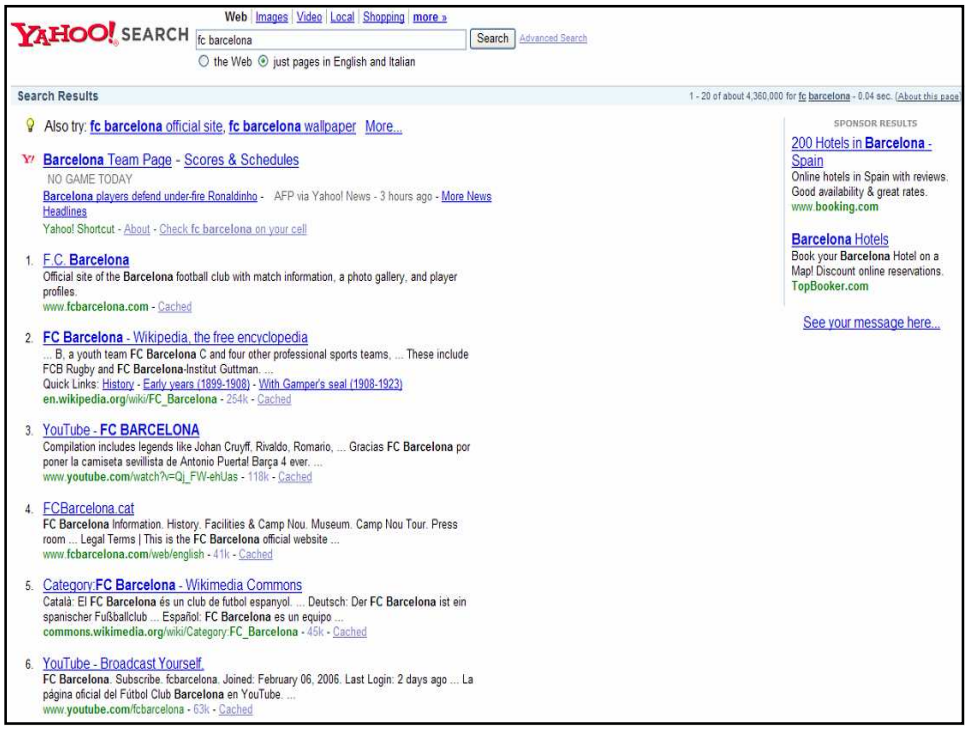
- SEs make possible
  - aggregation of interests
  - unlimited selection (Amazon, Netflix,...)

Incentives for specialized niche players

The biggest money is in the smallest sales !!

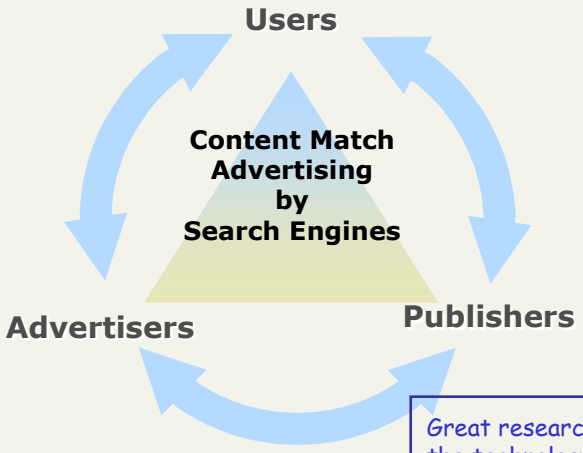



Visited Pages
Clicked Banner
Web Searches
Clicks on Search Results
CONTENT CONSUMED
ADS CLICKED
SEARCH CLICKS
Web
usage data !!!
YAHOO!


YAHOO! SEARCH
Web | Images | Video | Local | Shopping | more »
fc barcelona
Search
Advanced Search
the Web just pages in English and Italian
Search Results
Also try: fc barcelona official site, fc barcelona wallpaper More...
Barcelona Team Page - Scores & Schedules
NO GAME TODAY
Barcelona players defend under-fire Ronaldinho - AFP via Yahoo! News - 3 hours ago - More News Headlines
Yahoo! Shortcut - About - Check fc barcelona on your cell
1. F.C. Barcelona
Official site of the Barcelona football club with match information, a photo gallery, and player profiles.
www.fcbarcelona.com - Cached
2. FC Barcelona - Wikipedia, the free encyclopedia
... B, a youth team FC Barcelona C and four other professional sports teams, ... These include FCB Rugby and FC Barcelona-Institut Guttman. ...
Quick Links: History - Early years (1899-1908) - With Gamper's seal (1908-1923)
en.wikipedia.org/wiki/FC_Barcelona - 254k - Cached
3. YouTube - FC BARCELONA
Compilation includes legends like Johan Cruyff, Rivaldo, Romario, ... Gracias FC Barcelona por poner la camiseta sevillista de Antonio Puerta! Barça 4 ever. ...
www.youtube.com/watch?v=Qj_FW-ehUas - 116k - Cached
4. FCBarcelona.cat
FC Barcelona Information. History. Facilities & Camp Nou. Museum. Camp Nou Tour. Press room ... Legal Terms | This is the FC Barcelona official website ...
www.fcbarcelona.com/web/english - 41k - Cached
5. Category:FC Barcelona - Wikimedia Commons
Català: El FC Barcelona és un club de futbol espanyol. ... Deutsch: Der FC Barcelona ist ein spanischer Fußballclub ... Español: FC Barcelona es un equipo ...
commons.wikimedia.org/wiki/Category:FC_Barcelona - 45k - Cached
6. YouTube - Broadcast Yourself.
FC Barcelona. Subscribe. fcbarcelona. Joined: February 06, 2006. Last Login: 2 days ago ... La página oficial del Fútbol Club Barcelona en YouTube. ...
www.youtube.com/fcbarcelona - 63k - Cached
SPONSOR RESULTS
200 Hotels in Barcelona - Spain
Online hotels in Spain with reviews. Good availability & great rates.
www.booking.com
Barcelona Hotels
Book your Barcelona Hotel on a Map! Discount online reservations.
TopBooker.com
See your message here...

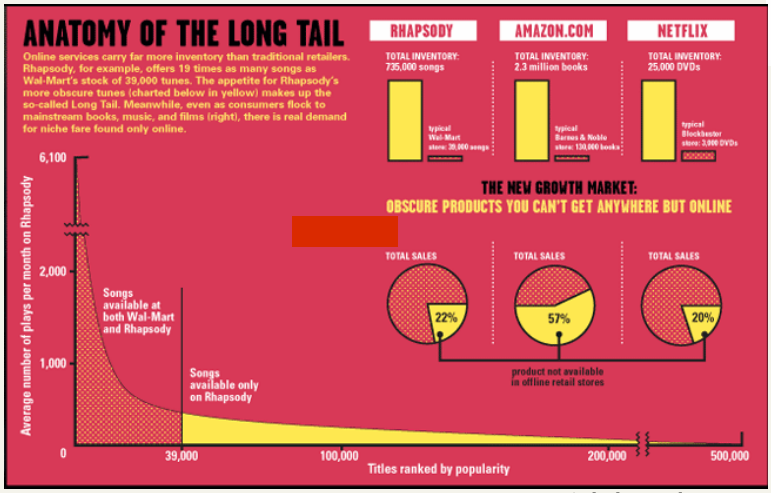
# Ad Industry

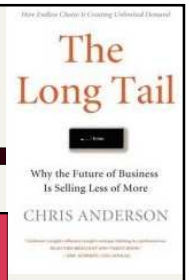
At Yahoo! we are investigating a "Theory of Information Supply"




Great research opportunities: we can build the technology to solve the business problem, and we can change the business problem to make it solvable.

# The long tail




(Chris Anderson, Wired, Oct 2004)

## A new game

Similar to web searching, but: Ad-DB is smaller, Ad-items are small pages, ranking depends on clicks

- For advertisers:
  - What words to buy, how much to pay
  - SPAM is an economic activity
- For search engines owners:
  - How to price the words
  - Find the right Ad
  - Keyword suggestion, geo-coding, business control, language restriction, proper Ad display

## Web Algorithmics
(The power of Algorithms)

Recommendation Systems
(not only Web Search…)

# Recommendations

- We have a list of restaurants
  - with ↑ and ↓ ratings for *some*

| | Brahma Bull | Spaghetti House | Mango | Il Fornaio | Zao | Ming's | Ramona's | Straits | Homma's |
|---|---|---|---|---|---|---|---|---|---|
| Alice | | Yes | No | Yes | | | | No | |
| Bob | | Yes | | | | No | | No | |
| Cindy | | | | Yes | No | | | No | |
| Dave | No | | | No | Yes | Yes | | | Yes |
| Estie | | | | No | Yes | Yes | | Yes | |
| Fred | No | | | | | | No | | |

Which restaurant(s) should I recommend to Dave?

# Basic Algorithm

- Recommend the **most popular** restaurants
  - say # positive votes minus # negative votes

| | Brahma Bull | Spaghetti House | Mango | Il Fornaio | Zao | Ming's | Ramona's | Straits | Homma's |
|---|---|---|---|---|---|---|---|---|---|
| Alice | | 1 | -1 | 1 | | | | -1 | |
| Bob | | 1 | | | | -1 | | -1 | |
| Cindy | | | | 1 | -1 | | | -1 | |
| Dave | -1 | | | -1 | 1 | 1 | | | 1 |
| Estie | | | | -1 | 1 | 1 | | 1 | |
| Fred | -1 | | | | | | -1 | | |

- What if Dave does not like Spaghetti?

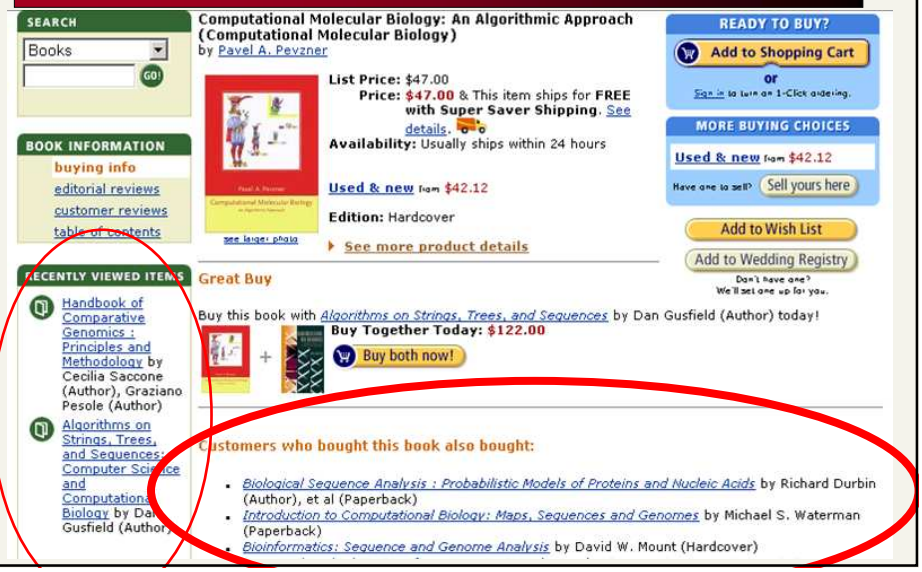
# Smart Algorithm

- **Basic idea**: find the person "most similar" to Dave according to **cosine-similarity**, and then recommend something this person likes.

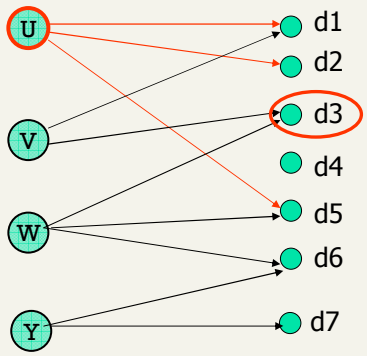
| | Brahma Bull | Spaghetti House | Mango | Il Fornaio | Zao | Ming's | Ramona's | Straits | Homma's |
|---|---|---|---|---|---|---|---|---|---|
| Alice | | 1 | -1 | 1 | | | | -1 | |
| Bob | | 1 | | | | -1 | | -1 | |
| Cindy | | | | 1 | -1 | | | -1 | |
| Dave | -1 | | | -1 | 1 | 1 | | | 1 |
| Estie | | | | -1 | 1 | 1 | | 1 | |
| Fred | -1 | | | | | | -1 | | |

☝ Do you want to rely on one person's opinions?

# Item2Item collaborative filtering



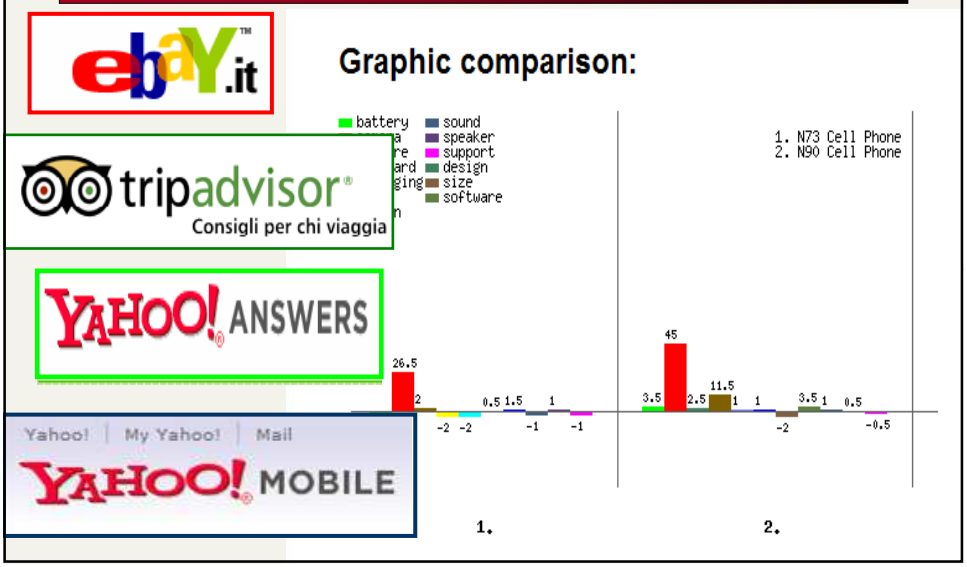
# Main idea




What do we suggest to U ?

Suggest the item(s) which are popular among the users that bought the same things as U

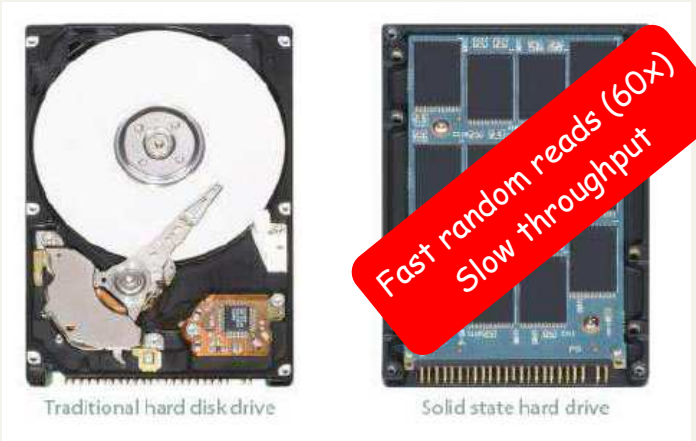
# An interesting issue



## Web Algorithmics
(The power of Algorithms)

My main interest…

## Solid state hard drives




512Bytes → 4 Kb  16Kb → 512 Kb

## Some tests on PC

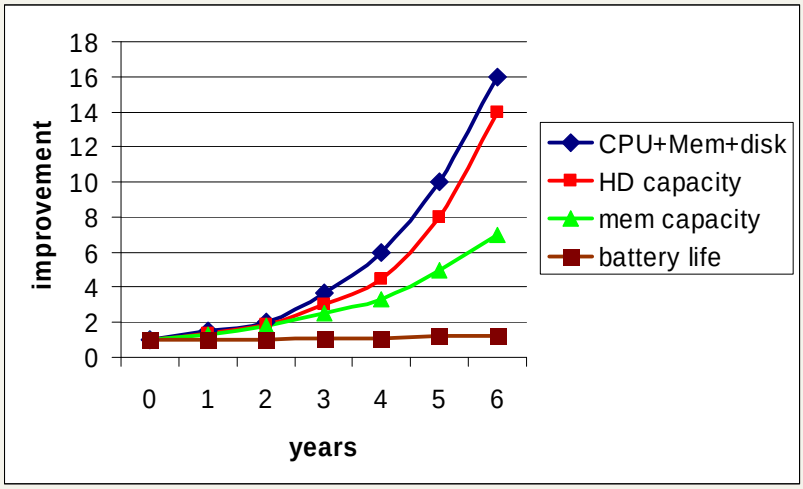
[Ajwani et al, 2008]

| Experiment | Time/ele SSD | Time/ele HD | |
|---|---|---|---|
| Generating & writting $2^{30}$ random doubles | 0.4 $\mu$ sec | 0.08 $\mu$ sec | x5 |
| Scanning (Read) | 0.30 $\mu$ sec | 0.09 $\mu$ sec | x3 |
| Sorting | 1.6 $\mu$ sec | 0.44 $\mu$ sec | x4 |
| Random read | 0.4 msec | 5.5 msec | x14 |
| Binary Search | 1.7 msec | 12.4 msec | X7 (some locality) |

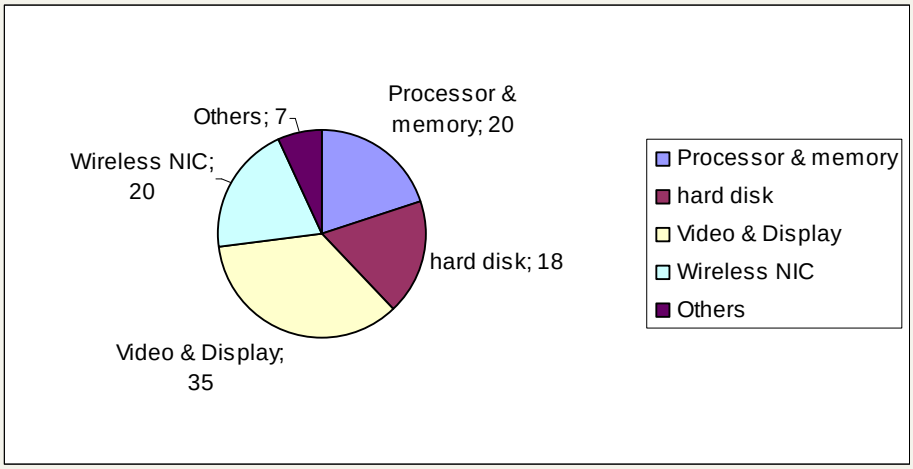
List-ranking a randomly stored list of $2^{30}$ element list of long integers by naïvely following pointers takes minutes in RAM, but days with flash.

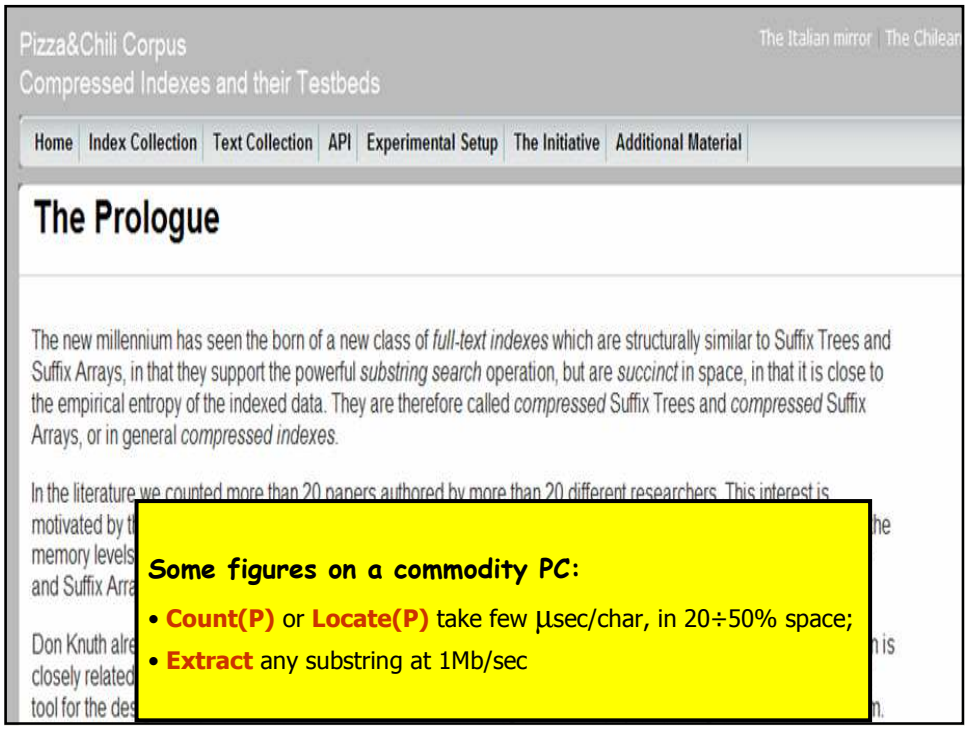
Need to distinguish between read/write

## Energy *vs* capacity *vs* performance



# Energy efficiency



# Take-away message

## Data is precious


**Large volumes, occur in various contexts:**

- Telephone calls
- Bills: supermarket, shops, …
- BioInformatics: DNA, 3D-mapping,…
- Web sites: navigation and search logs
- Mobile phones: location, search, browse

…but more data does **not** mean more information.

You need proper algorithms !