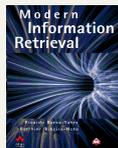


# Web Algorithmics (The power of algorithms)

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!** Web Images Video Local Shopping more

Search:  **Web Search** the Web Italy only

My Yahoo! My Mail Page Options

**Answers**

**Autos**

**Finance**

**Games**

**Groups**

**HotJobs**

**Maps**

**Mobile Web**

**Movies**

**Music**

**Personals**

**Real Estate**

**Shopping**

**Sports**

**Tech**

**Travel**

**TV**

**Yellow Pages**

**Bix**

**More Yahoo! Services**

**Small Business**

- Get a Web Site
- Domain Names
- Sell Online
- Search Ads

**Featured Services**

**Featured** Entertainment Sports Video

Sep 21, 2007



### 400 richest Americans

For the first time, it takes more than a billion to make Forbes' wealthiest list. **Two new members of top 10**

- Top 22
- How to make a million
- World's richest man's new charity

Forbes' list of the 400 richest Americans

Quick meal ideas for breakfast on the go

The best places in the world to live

Sheen and Richards' custody battle turns ugly

**More Featured**

**In the News** World Local Finance

As of 4:04 p.m.

- Bombing and heavy fighting kills 82 in Afghanistan
- Iraq convoys under Blackwater protection resume
- Israel urged to turn over Arab areas for peace deal
- Two students shot at Delaware State University
- Reported Florida tornado damages 50 homes
- Doctors to separate 2-year-old conjoined twins
- Philadelphia artist uses worms to create abstract works

Photos Suspension Search Damage Science

**More: News Election '08: Candidate Mashup**

Markets: Dow: +0.5% Nasdaq: +0.5% Sponsored by: **Scottrade**

Stock Quotes:  **Go**

**Marketplace**

Popular action games

Turtle Odyssey 2, Lego Fever. Download free trials of the best action games at Yahoo! Games. Download now.

Bring movie night home with a new home theater - Pick the right setup

Hi, Paolo Sign Out

Mail Messenger Radio

Weather 72°F Local Horoscopes

**Yahoo! Travel** Flights Cars Deals Vacations Hotels

**Plan Your Getaway**

Today's Top Deals Vacation Packages\* Huge Cruise Savings\* Hotel Deals Worldwide

\*Info on taxes and fees

Search Travel: Enter City Name **Go**

**Be a Better Movie Date**

Watch movie trailers and clips

Good Luck Chuck Resident Evil... Sydney White Jesse James

**Pulse - What Yahoos Are Into**

**All Stars: Most Viewed Cars**

Land Rover LR2

Nissan Altima Coupe

Rolls-Royce Phantom

Lotus Elise

Nissan Altima

BMW X5

Toyota Corolla

Ford Shelby GT500

**Ask** .com

Web Images City News More

beatles  **Q**

Advanced

**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

**beatles** Showing 1-10 of 9,787,000 MyStuff Options



**The Beatles** | Save

One of the biggest musical acts in history, The Beatles were John Lennon (guitar), George Harrison (guitar), Paul McCartney (bass) and Ringo Starr (drums). Lennon and McCartney began playing together in The Quarrymen in 1957, Harrison joined later that year. Before... [More »](#)

Search For: [Ringtones](#)

Go To: [Music](#) | [Encyclopedia](#)

Other matches:  **Go**

**THE BEATLES**

Detailed history with information on their music, movies, news, and latest projects. Images, related links, and a showcase for their albums.

[www.thebeatles.com/](http://www.thebeatles.com/) - Cached

**THE BEATLES**

The official website of The Beatles...

[www.beatles.com/](http://www.beatles.com/) - Cached

**Songs, Pictures, and Stories of The Beatles**

Beatles website for collectors and fans featuring information and values on thousands of Beatles rarities...

[www.rarebeatles.com/](http://www.rarebeatles.com/) - Cached

**Frank's meagre Beatles page**

Links to articles and annotated discographies of Beatles recordings.

[members.aol.com/egweimi/bt11.htm](http://members.aol.com/egweimi/bt11.htm) - Cached

**Images**



**Popular Tracks**

- Beatles
- Let It Be
- Hey Jude

Source: [Link](#)

**Encyclopedia**

**The Beatles**

The Beatles were an English rock band from Liverpool whose members

**YAHOO! LOCAL** Welcome, paolo.ferragina Sign Out 70°F

City Guide | My Local | Directory

hotel Manhattan, NY Search

Manhattan City Guide > Travel & Lodging > Hotels & Lodging > Hotels & Motels

CATEGORY SPONSORS

- SOHOTEL** - Clean, Courteous, Moderate.  
 (866) 629-4995, 341 Broome St, New York, NY  
[Get directions](#) [www.sohotel-nv.com](#)
- Gramercy Park Hotel** - Exclusive Packages & Lowest Rates - Book Direct  
 (212) 920-3300, 2 Lexington Ave, New York, NY  
[Get directions](#) [www.gramercyparkhotel.com](#)
- Hotel Low Rates**  
 Low rate guarantee at discount hotels. [Hotels.com](#).  
[www.hotels.com](#)

Events for hotel near Manhattan, NY

- SEP 25** Wes Anderson's Hotel Chevalier - Special Screening (2 people) - Apple Store SoHo
- OCT 24** The Shins (10 people) - Terminal 5
- OCT 21** Jump Start Your Celebrity Makeup Hair & Fashion Styling Career for Print, Video, Film & TV (6 people) - MAC Pro Store

[See more events from Upcoming >](#)

Results 1-10 out of 1830 total ([About these results](#))

[Narrow your search](#) by category, rating, and more.

Sorted by: **top results** | distance [Print results](#)

**The Plaza** ★★★★★ (314) 1.24 mi

(212) 588-8000  
 768 5th Ave, New York, NY  
[Get Directions](#)  
[www.theplazaresidences.com/](#)  
 Neil Simon comedy. This is one of New York's

**Zoom in and search the map**

SPONSOR RESULTS

**Hotel Reservation at Booking.com**  
 Compare all hotels online, book now and save up to 75%.  
[www.booking.com/hotel-reservation](#)

**Belize hotel-Belize City hotel Belize**  
 Belize hotel-Belize City hotel-Bachelor Inn-Near University of...  
[belize-hotel.01come.com](#)

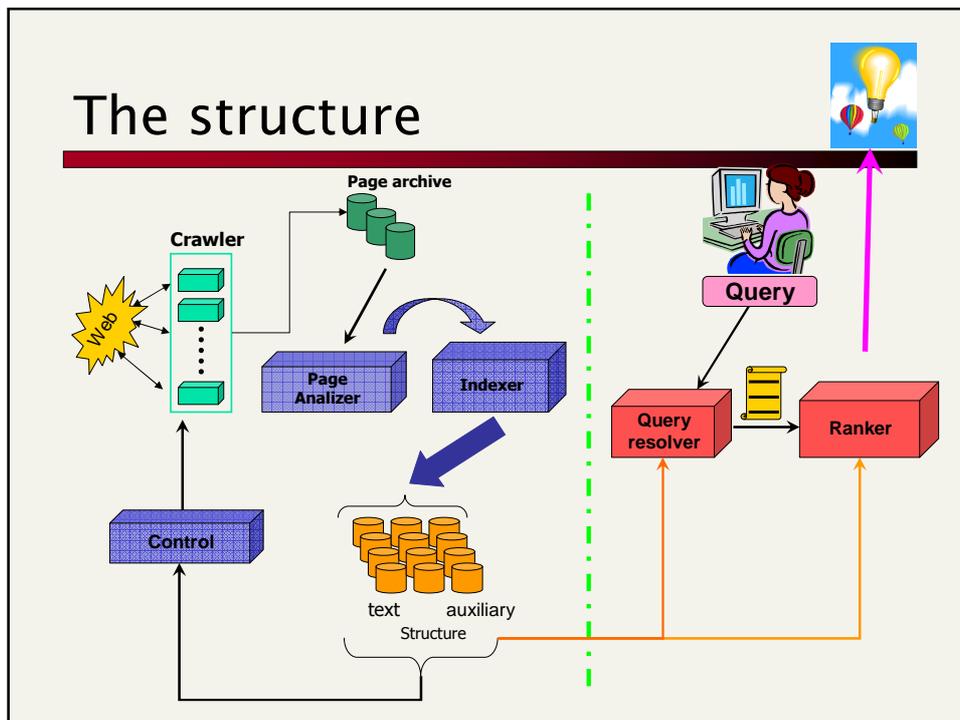
**Hotel**

This is a search engine!!!



# Web Algorithmics (The power of algorithms)

## The structure of a Search Engine





Search the site with google

Main Wiki Lucene 2.3.1 Documentation Last Published: 02/24/2008 02:08:06

- ~ About
  - Overview
  - Features
  - Powered by Lucene
  - Who We Are
- Documentation
- Resources
- Related Projects

built with Apache Forrest

## Apache Lucene - Overview

[PDF](#)

- [Apache Lucene](#)
- [Lucene News](#)
  - [23 February 2008 - Lucene Java 2.3.1 available](#)
  - [24 January 2008 - Lucene Java 2.3.0 available](#)
  - [23 January 2008 - Lucene at ApacheCon Europe](#)
  - [24 December 2007 - Nightly Snapshots available in the Apache Maven Snapshot Repository](#)
  - [26 August 2007 - Lucene at ApacheCon Atlanta](#)
  - [19 June 2007 - Release 2.2 available](#)
  - [18 February 2007 - Lucene at ApacheCon Europe](#)
  - [17 February 2007 - Release 2.1 available](#)
  - [3 January 2007 - Nightly Source builds available](#)

---

### Apache Lucene

Apache Lucene is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform.

Apache Lucene is an open source project available for [free download](#). Please use the links on the left to access Lucene.

## Web Algorithmics (The power of algorithms)

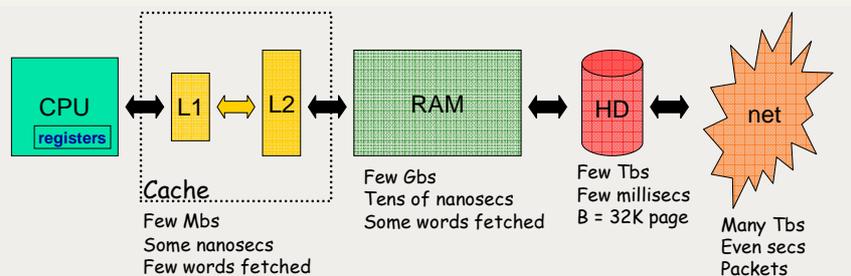
Algorithmic Issues:  
I/O and space

## 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** ?
  - $n_1 = k * t$ ,

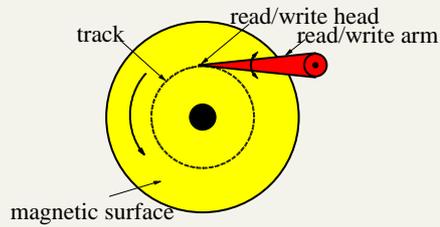
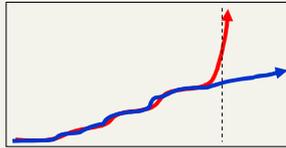
## Algorithm Inadequacy

- 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

IBM

## 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  $\approx 10$  Gbytes
  - # docs  $\approx 4 * 10^6$
  - #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  $\approx 4M$

	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

#terms  $> 1M$

Space  $\approx 4Tb$  !

1 if **play** contains  
**word**, 0 otherwise

## 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<sup>5</sup> ÷ 10<sup>6</sup> (Hennessy-Patterson)]

If  $N=(1+f)M$ , then the  $\Delta$ -avg cost per step is:

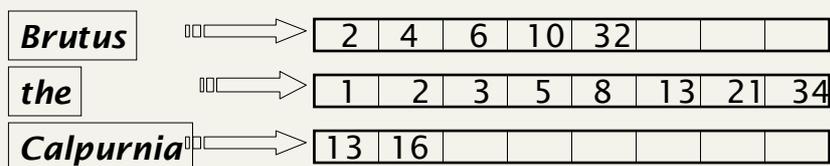
$$C * p * f / (1+f)$$

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

If we fetch  $B \approx 4\text{Kb}$  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

## $\gamma$ -code for integer encoding

0000.....0	x in binary
Length-1	

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

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

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

## It is a prefix-free encoding...

0001000001100110000011101100111

8

6

3

59

7

## Variable-byte codes

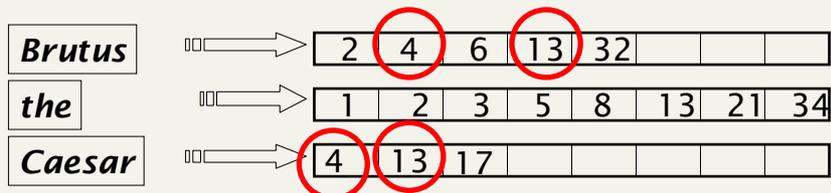
- Wish to get **very fast** (de)compress → **byte-align**

e.g.,  $v=2^{14}+1 \rightarrow \text{binary}(v) = 100000000000001$

1 0000001 1 0000000 0 0000001

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

## Query processing

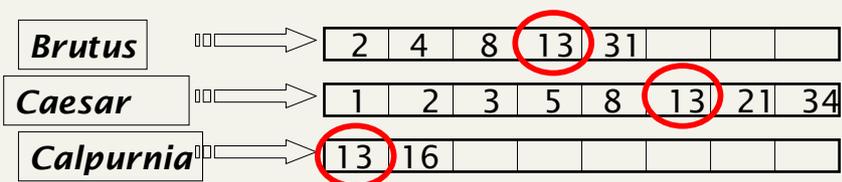


1) Retrieve all pages matching the query

## Some optimization

Best order for query processing ?

- Shorter lists first...

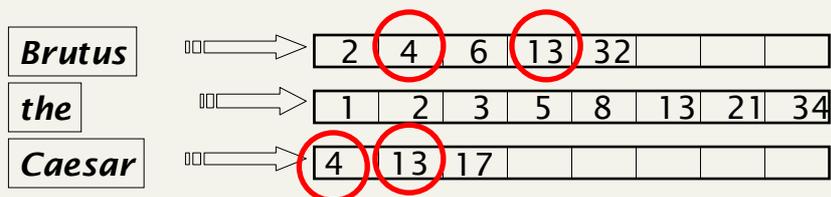


Query: **Brutus AND Calpurnia AND Caesar**

## 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 !

Google data compression Search [Advanced Search](#) [Preferences](#)

Web Results 1 - 10 of about 5,980,000 for [data compression](#). (0.20)

**Data compression - Wikipedia, the free encyclopedia**  
 In computer science and information theory, **data compression** or source coding is the process of encoding information using fewer bits (or other ...  
[en.wikipedia.org/wiki/Data\\_compression](http://en.wikipedia.org/wiki/Data_compression) - 78k - [Cached](#) - [Similar pages](#)

**Data Compression**  
 This paper surveys a variety of **data compression** methods spanning almost forty years of research, from the work of Shannon, Fano and Huffman in the late ...  
[www.ics.uci.edu/~dan/pubs/DataCompression.html](http://www.ics.uci.edu/~dan/pubs/DataCompression.html) - 10k - [Cached](#) - [Similar pages](#)

**Data-Compression.com**  
 A website devoted to the principles and practice of **data compression**.  
[www.data-compression.com/](http://www.data-compression.com/) - 8k - [Cached](#) - [Similar pages](#)

**Interactive Data Compression Tutor**  
 This 'Interactive **Data Compression Tutor**' is a web-based teaching aid for **data compression**. It includes information about the fundamental principles and ...  
[www.eee.bham.ac.uk/woolleysi/A117/body0.htm](http://www.eee.bham.ac.uk/woolleysi/A117/body0.htm) - 5k - [Cached](#) - [Similar pages](#)

**Sponsored Links**

**Tension Data**  
 Leading manufacturer of safe load & line tension monitoring instruments  
[www.RobWay.com.au](http://www.RobWay.com.au)

**Data Compression**  
 Index & compress with IRISPdf.  
 High volume OCR solution!  
[www.irislink.com](http://www.irislink.com)

**Data Compression**  
 Get up to 1000% more WAN compression with Expand  
[MyExpandNetworks.com](http://MyExpandNetworks.com)

**Data Compression**  
 Decode PSTN & IP Information. [Data](#)

# An interesting issue

## Compressed DB of Web pages → Snippets !!

- ❖ Gbs/Tbs of data
- ❖ Data are heterogeneous
- ❖ Support (random)

### Two main (compression)

- ❖ Gzip etc.
- ❖ Huffword

Whatever is your idea,  
it must be SCALABLE

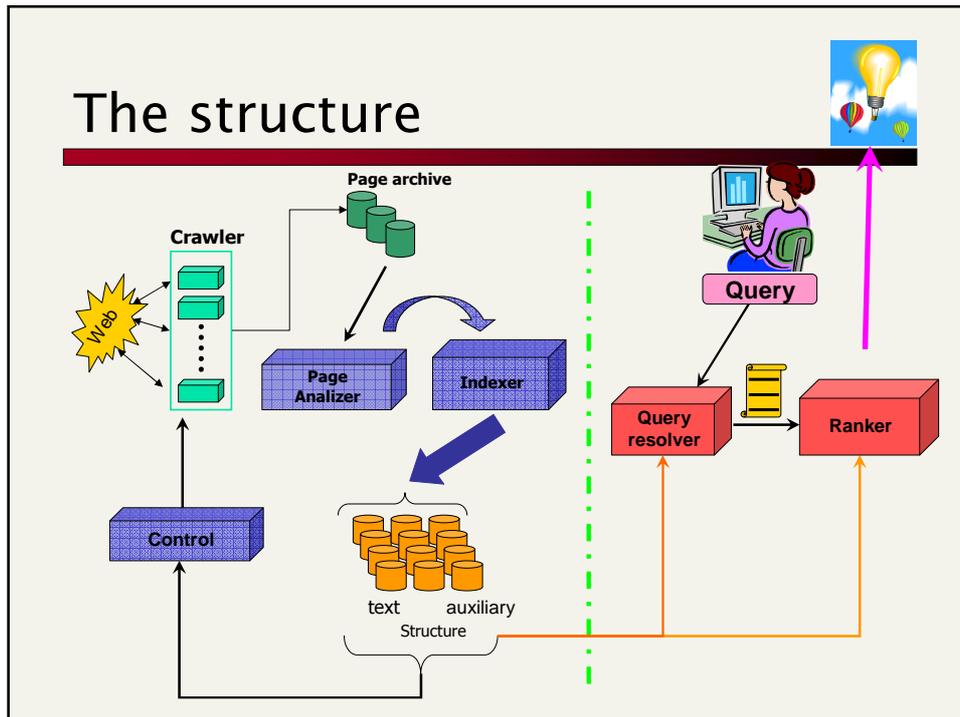
### Various issues

1. Clustering
2. Parsing = Dictionary selection
3. (Optimal) Model in limited space

### On 47Gb of UK-pages [UniMi]

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

# The structure



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

Comparing Google and Yahoo! Search results 1 - 100 for "mousetrap".

mousetrap      compare 'em

Google

touch the dots!

YAHOO!

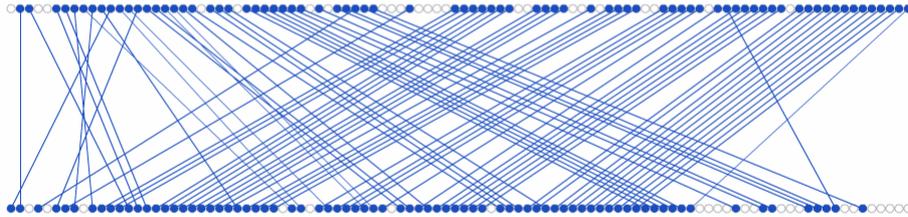
## Ranking: Google vs Google.cn

Comparing Google and Google.cn results 1 - 100 for "comunism"

comunism

Google

<http://www.marxists.org/romana/dictionar/c/Comunism.htm>



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  $d = \#occ_t / |d|$

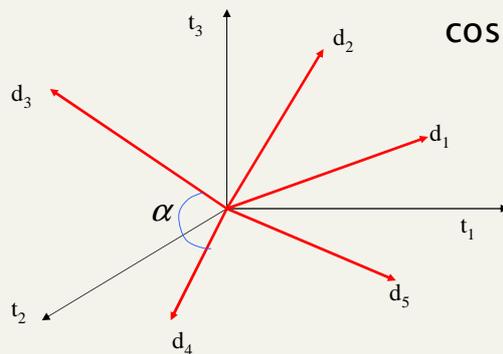
$idf_t = \log\left(\frac{n}{n_t}\right)$  where  $n_t = \#docs$  containing term  $t$   
 $n = \#docs$  in the indexed collection

	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				

Vector Space model

## A graphical example

Easy to Spam



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

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.

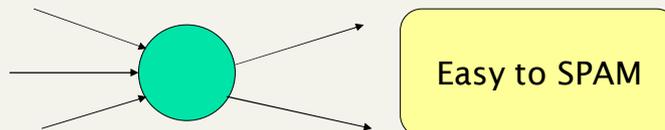
Now SE use tf-idf PLUS PageRank (PLUS other weights)

Web Algorithmics  
(The power of Algorithms)

Link-based Ranking  
(2° generation)

## Query-independent ordering

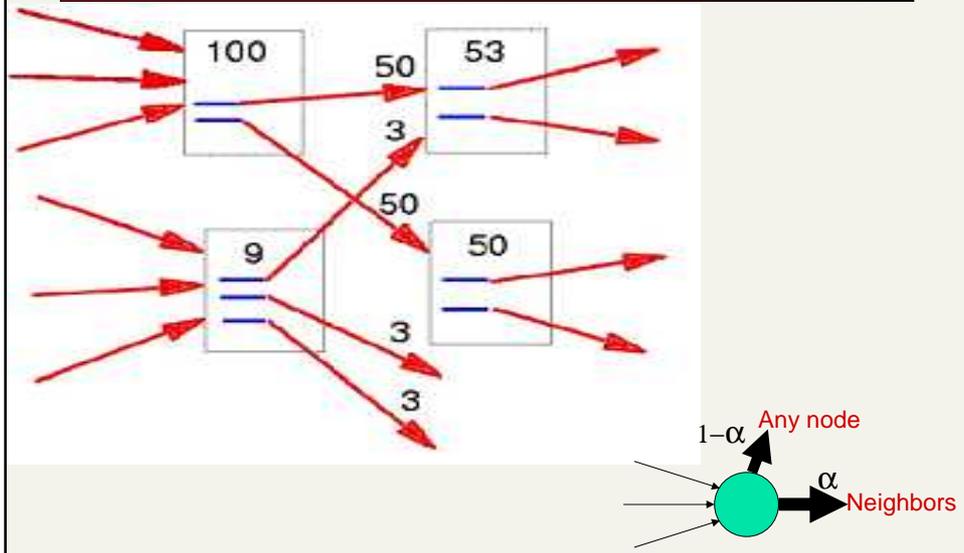
- First generation: using link counts as simple measures of popularity.
  - Undirected popularity:
    - Each page gets a score given by the number of in-links plus the number of out-links (es.  $3+2=5$ ).
  - Directed popularity:
    - Score of a page = number of its in-links (es. 3).



## 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

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

**B(i)** : set of pages linking to i.

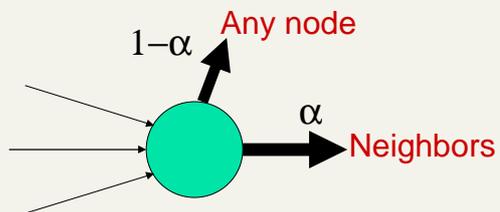
**#out(j)** : number of outgoing links from j.

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

## 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

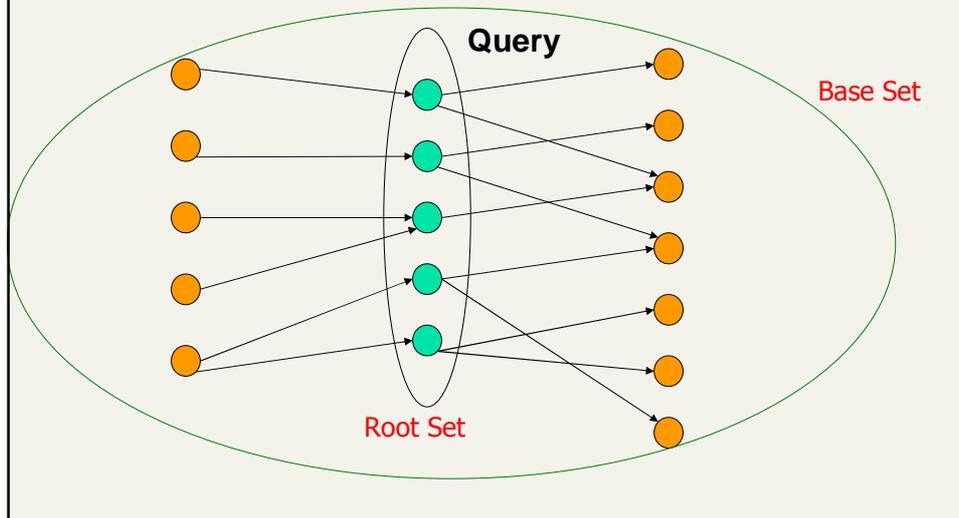
- Preprocessing:
  - Given graph of links, build matrix **P**
  - Compute its **principal eigenvector r**
  - $r[i]$  is the pagerank of page  $i$

*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*

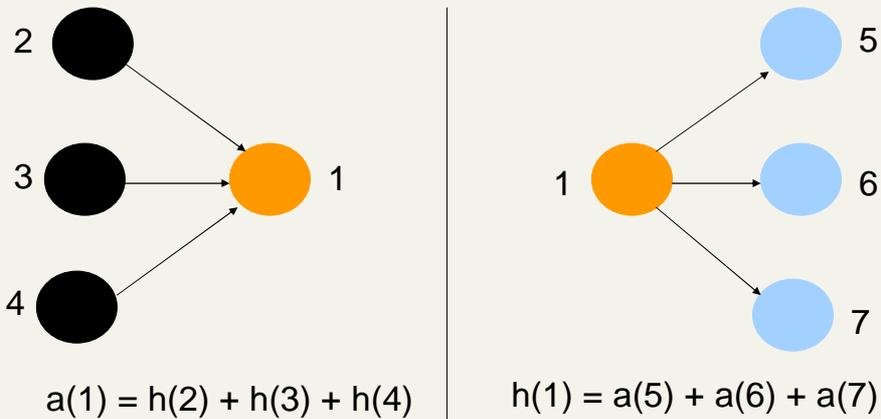
## HITS: Hypertext Induced Topic Search



## 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



## HITS: Link Analysis Computation

$$\left. \begin{array}{l} a = A^T h \\ h = Aa \end{array} \right\} \Rightarrow \begin{array}{l} a = A^T Aa \\ h = AA^T h \end{array}$$

Where

a: Vector of Authority's scores

h: Vector of Hub's scores.

A: Adjacency matrix in which  $a_{ij} = 1$  if  $i \rightarrow j$

**Thus,  $h$  is an eigenvector of  $AA^T$   
 $a$  is an eigenvector of  $A^T A$**

## Weighting links

---

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

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

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



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

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

## Web Algorithmics (The power of algorithms)

Clustering of Search Results  
(a different approach...)

about | products | solutions | press | partners | support

Vivísimo  
search.vivísimo.com

mousetrap the Web Search [Advanced Search](#) [Help](#)

NEW read the [latest gossip](#) at [Clusty.com](#)

**Clustered Results** Top 194 results of at least 1,309,763 retrieved for the query **mousetrap** (Details)

- ▶ [mousetrap](#) (198)
- ▶ [Mousetrap Cars](#) (44)
- ▶ [Better Mousetrap](#) (30)
- ▶ [Photos](#) (19)
- ▶ [Theatre](#) (14)
- ▶ [Machines](#) (12)
- ▶ [Christie, Agatha](#) (13)
- ▶ [Reviews](#) (8)
- ▶ [Humane](#) (8)
- ▶ [Children](#) (8)
- ▶ [Sales](#) (4)
- ▼ [More](#)

Find in clusters:  
Enter Keywords

**The Mousetrap** Sponsored Link  
Buy Tickets to all top London shows Buy **Mousetrap** Tickets Online!  
[www.LondonTheatreBoxOffice.net](#) - Sponsored Listings 1

**Hotels near The Mousetrap** Sponsored Link  
75% off London Hotels. Find same deal for less & we'll pay you £100.  
[www.LondonTown.com](#) - Sponsored Listings 2

1. [Mousetrap cars and mouse trap vehicle kits, books, and plans](#) [new window] [frame] [cache] [preview] [clusters]  
Offers a book about **mousetrap** powered vehicles, as well as kits, plans and parts. Also has information and construction hints on site.  
[www.docfizzix.com](#) - Gigablast 2, Ask 3, Open Directory 8, Live 10
2. [Concord MouseTrap](#) [new window] [frame] [cache] [preview] [clusters]  
Online retailer in Concord, MA, where WFF are made. Monthly Squeak Newsletter, photos of current and retired pieces, and discontinued pieces for sale.  
[www.concordmousetrap.com](#) - Gigablast 5, Open Directory 9, Live 14, Open Directory 21
3. [Mousetrap Backpackers - Paihia, Bay of Islands, NZ](#) [new window] [frame] [cache] [preview] [clusters]  
Accommodates only 25 people, and is a character house for character people. The hostel backs onto native bush, and has a large lovely garden in front.  
[www.mousetrap.co.nz](#) - Open Directory 5, Live 18, Gigablast 19, Ask 71

## 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

**Personalized SnakeT**  
 SNippet Aggregation for Knowledge ExTraction ALPHA

Great tools... both the compar engine. Your (clusterin  
 I tried it a bit, looks quite nice say how much Google wo  
[Look What people...](#)

Search

Web [\[select all\]](#) [\[select none\]](#)

Google  YAHOO!  altavista  alltheweb  
 TEOMA  loooksmart  overture  msn  
 About  mozDex  AMERICA Online  findwhat  
 GIGABLAST  esporting  A9  Bloglines  
 entireweb  ixquick  mamma  
 Netscape  WiseNut

Literature  A9  SCIRUS for scientific information only

News  Google News

## 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

The screenshot displays the SnakeT search interface. On the left, a 'Clusters' sidebar shows a hierarchical tree of search results. The 'Java' cluster is expanded, showing sub-clusters like 'Technology', 'Programming', 'Tutorials', 'Free', 'Training', 'Developers', 'Java Books', 'Features Java', 'Coffee', 'Site For Java', 'Games', 'Java Index', 'Java Environment', 'Java Forums', and 'Virtual Machine'. The 'Tutorials' cluster is highlighted in yellow. On the right, the 'Search' results pane shows a list of search results for 'The Java Tutorial'. The results include:

- [The Java Tutorial](#)  
... Sun Microsystems. Developers Home Products Technologies Java Technology Learning Tutorial Tutorial. ...  
[google:2]
- [Java\(TM\) Boutique - Programming Tutorials, Reviews and Downloads](#)  
The Java Boutique is a collection of java applets, games, scripts, and tutorials. Learn programming as also find news about java and jni. ... Programming languages have evolved from machine language t what's wrong and why it's necessary. The Java Memory Model Explained ...  
[altavista:2 | google:5 | man:4 | looksmart:6 | yahoo:4]
- [JavaScript Kit- Comprehensive JavaScript, DHTML, CSS tutorials and ...](#)  
JavaScript Kit Formerly Website Abstraction, Click Here. ...  
[google:16]
- [Java Programming Resources -- Java, Java, and more Java](#)  
Java programming resources: FAQs, tutorials, compiler and browser download sites, documentation,  
[google:18]
- [Molecular Expressions: Science, Optics and You - Secret Worlds ...](#)  
... protons. Interactive Java Tutorial, ATTENTION. ... functioning property. Please install this softwa  
[google:24]
- [Welcome to Freewarejava.com, the place to find free Java applets ...](#)

# BioPromptBox [Ferragina et al, Bioinformatics 2007]

the **BioPrompt-box** (v. 0.8.5)

glutamate UNIPROT Search! BLAST

Set a query and choose a data bank where to search.

Views

GO Upward Paths Terms Taxonomy Organisms Keywords

Filter Selection Execute

Open All Close All

"glutamate" (100)

- molecular function (89)
  - glutamate receptor activity (42)
    - ionotropic glutamate receptor activity
    - metabotropic glutamate, GABA-B-like (14)
  - protein binding (27)
  - metabotropic glutamate, GABA-B-like rec
  - L-glutamate transporter activity (9)
  - ion channel activity (8)
  - carrier activity (4)
  - glutamate-cysteine ligase activity (4)
  - glutamate decarboxylase activity (2)

1.  **Glutamate racemase** <sup>id</sup>

Entry Name: MURI\_CARST

Organism: Carnobacterium sp. (strain St2)

Comment(s):

- Provides the (R)-glutamate required for cell wall biosynthesis
- L-glutamate = D-glutamate
- more...

Citation(s):

- Proteins from cold-adapted bacteria: evolutionary and structural relationships with mesophilic and thermophilic counterparts.**

Actions on this entry: [Keywords](#) | [BLAST it!](#) | [Highlight](#) | [Just this](#)

Score: 100.0%

2.  **Glutamate racemase** <sup>id</sup>

Entry Name: MURI\_AQUAE

# An interesting issue

Yahoo! My Yahoo! Mail

**YAHOO! MOBILE** Search: Web S

Home Mobile Services Developers News & Partners

Developer Beta

**Maximum Reach. Minimum Effort.**

Yahoo! introduces Mobile Widgets—the easiest way to reach the most mobile consumers today

**Building a Widget**

Create

Learn about the different types of Mobile Widgets

**Yahoo! Mobile Developer Platform**

Be one of the first to preview our new open platform built to maximize your reach—across hundreds of mobile devices. The Yahoo! Mobile Developer Platform will empower developers like you to:

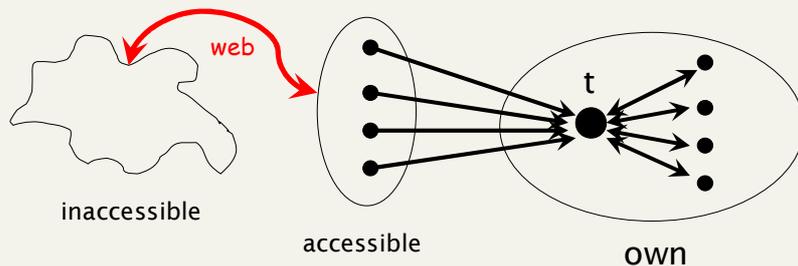
**Get Started**

View SDK Now

# Web Algorithmics (The power of algorithms)

Web SPAM  
(the real competitor...)

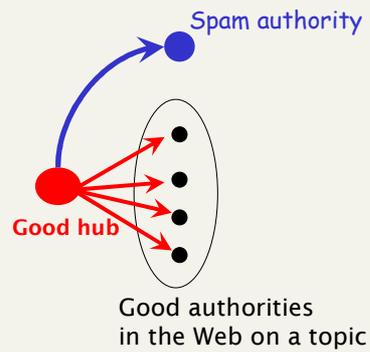
## Spamming PageRank



It is NOT easy to spam

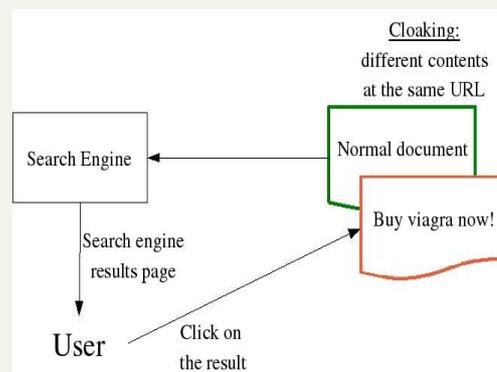
## Spamming HITS

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

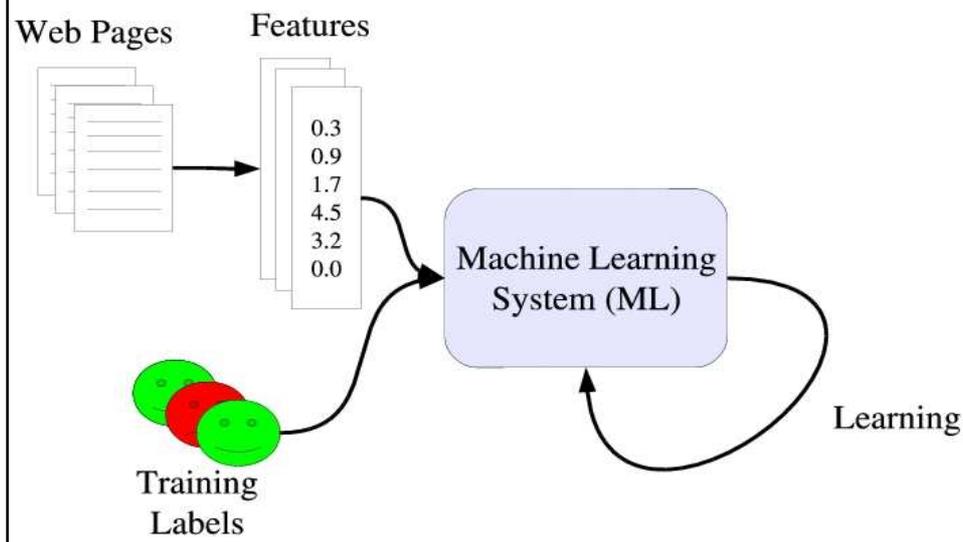
## Some examples of SPAM...

Made for advertising Search engine?

The screenshot shows a web browser window with a search engine results page. The search query is "sports book". The results are dominated by advertisements and search engine-related content, illustrating spam. The page includes a search bar, a list of top searches, and a list of top web results. The top web results are:

- Place Your Bet with #1 Sports Betting Site Online Kentucky Derby, NBA, NFL, NHL, and all other sports betting and odds. Place a full race sportsbook in North America. <http://www.sportsinteraction.com>
- AutoUp GamblingLinks.com - Safe Online Casinos Link to safe and secure online casino gambling and sports betting including reviews. <http://gamblinglinks.com>
- Free Casino Bonuses. Links To the Best Casinos Get \$20 - \$500 in Free Chips. Most popular casino games w/ great graphics. Play for free rules and strategy. Link to the Best Casinos. <http://www.fasttrotzash.net>
- AutoUp GamblingLinks.com - Safe Online Casinos

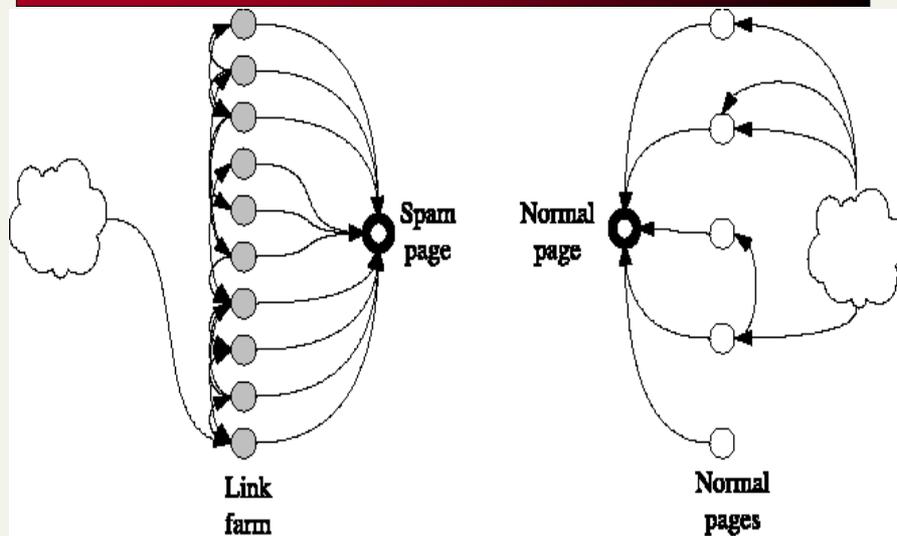
## Typical approach



## Content-based features

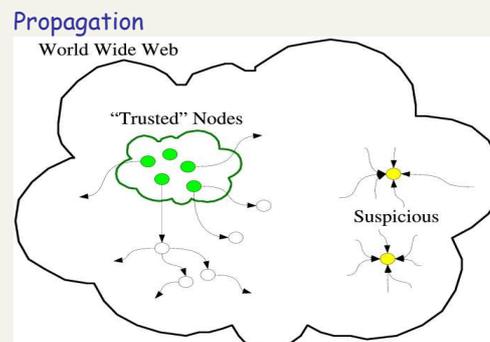
- number of out/in links
- number of words in the body/title
- avg word length
- fraction of anchor text or visible text
- compression rate

## 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



## An interesting issue

---

Pinter's graphlet vectors

- Characterize Web pages
- Characterize Countries

- Public spam collection
  - Labels for 6000000 pages
  - 2,725 hostnames
  - 3,106 automatically considered normal (.ac.uk, .sch.uk, .gov.uk, .med.uk, .nhs.uk or .police.uk)
  - <http://www.yr-bcn.es/webspam/>

Whatever is your idea,  
it must be SCALABLE

## 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





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

## Two approaches

- **Sponsored search:** Ads driven by search keywords

AdWords

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

AdSense

The screenshot shows a Yahoo! search results page for the query "hotel excelsior new york". The page includes a search bar, navigation links (Web, Images, Video, Local, Shopping, more...), and search results. A red box highlights the sponsored results section, which includes:

- Search Results for Excelsior Hotel NY**
  - [www.excelsiorhotelnyc.com](http://www.excelsiorhotelnyc.com) - Stay at a beautiful luxurious four-star New York City Upper west side.
  - Excelsior NY Deals**
    - [www.hotels.com](http://www.hotels.com) - Low rate guarantee at New York hotels. Book now at Hotels.com.
  - Save on Excelsior Hotel New York**
    - [travel.hotelsanddiscounts.com](http://travel.hotelsanddiscounts.com) - Save up to 70% off at Excelsior Hotel in New York. Check rates...
  - Excelsior Hotel Manhattan New York City**
    - [www.newyorkjourney.com](http://www.newyorkjourney.com) - Excelsior Hotel is located in the Manhattan Upper West Side area at...
- SPONSOR RESULTS**
  - A1 Discount Hotels - NY**
    - A1 Discount Hotels: No Booking Fees Great Customer Service Low Rates.
    - [www.A1-discount-hotels.com](http://www.A1-discount-hotels.com)
  - Excelsior Hotel New York**
    - Save on Excelsior Hotel in New York. View photos. Read descriptions...
    - [book.hotelreservations.com](http://book.hotelreservations.com)
  - New York Excelsior Hotel Reservations**
    - Excelsior Hotel New York City, NY Low Price Guarantee.
    - [www.HotelsForEveryone.com](http://www.HotelsForEveryone.com)
  - Hotel Excelsior**
    - Save up to 75% on this hotel. No reservation fee, cancellation fee...
    - [hotels.denmark-bookings.com](http://hotels.denmark-bookings.com)
  - The Excelsior Hotel, Hong Kong**
    - Official site. Luxury hotel in Victoria Bay, elegant rooms & suites.
    - [www.MandarinOriental.com](http://www.MandarinOriental.com)
  - Excelsior Hotel New York**
    - Find Low Hotel Rates in New York Hotel Deals from 100+ Sites.
    - [www.kayak.com](http://www.kayak.com)
  - Excelsior Hotel New York**
    - Search over 120 travel sites. Save up to 70% on New York hotels.
    - [Hotels.SideStep.com](http://Hotels.SideStep.com)
- Local Results**
  - Hotel Excelsior near New York - Local Results**
    - Excelsior Hotel** - \*\*\*\*\* (212) 362-9200 - 45 W 81st St, New York, NY - 4.47mi - [map](#)
    - Yahoo! Shortcut - [About](#) - [Send local info to your cell!](#)
  - Excelsior Hotel, New York City NY Reviews - Yahoo! Travel**
    - Excelsior Hotel, New York City, NY: Find the best deals, reviews, photos, rates, and availability for the Excelsior Hotel on Yahoo! Travel.
    - [travel.yahoo.com/p-hotel-362420-excelsior\\_hotel\\_manhattan-i](http://travel.yahoo.com/p-hotel-362420-excelsior_hotel_manhattan-i)
  - Excelsior Hotel**
    - The Excelsior Hotel is a beautiful luxury four-star landmark in Manhattan's ... Excelsior Hotel New York you will find luxury and elegance throughout the hotel ...
    - [www.excelsiorhotelnyc.com](http://www.excelsiorhotelnyc.com) - 14k - [Cached](#)
  - Excelsior Hotel New York, New York Hotels by CrsHotels**
    - The Excelsior Hotel is a beautiful luxurious landmark in New York City's Upper West Side. ... Excelsior Hotel New York. Hotels by Crshotels ...
    - [crshotels.com/search/...&source=Ysm-Excelsior-Hotel-New-York](http://crshotels.com/search/...&source=Ysm-Excelsior-Hotel-New-York)
  - Excelsior Hotel Hotel, New York, NY - Hotels at CheapTickets.com**
    - CheapTickets.com is the faster, cheaper way to book the Excelsior Hotel Hotel in New York. Rooms starting at
    - [www.cheaptickets.com/App/ViewSpecificHotelLP?masterId=10482](http://www.cheaptickets.com/App/ViewSpecificHotelLP?masterId=10482)
  - New York Hotels | Official Site Millennium Broadway Hotel New York**
    - [Hotels.NewYork.Hotels.com](http://Hotels.NewYork.Hotels.com)

**Green Garden Tips**  
www.greengardentips.com

- [About Us](#)
- [Contact Me](#)
- [Privacy Policy](#)

**Hydroponics Gardening - An Introduction To Hydroponics Gardening For Beginners (Part 3) Lighting**  
This, the third article in my series on the basics of hydroponics gardening, covers the differing li...

**Pruning the Backyard Grapevine**  
Jim Bruce, of Rist Canyon Vineyards, tells gardeners how to prune their backyard grapevines to balan...

Search our Articles  
  
 Titles  
 Titles & descriptions

Order articles by: [Submission date](#) | [Article title](#)

Go to page: [ 1 ] [ 2 ] [ 3 ] ... [ 31 ] [ 32 ] [ 33 ]

**How To Build Auto Lawn Sprinklers**  
At last, you will be able to Quickly and Easily Install the Automatic Lawn Sprinkler System You Have Always Wanted, in less time and for less money than any contractor you will ever find.

**Free Landscape Designs**  
Make Great Landscape Plans Fast. See Examples. Free Download!  
www.SmartDraw.com

Ads by Google

# How does it work ?

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

IR

Econ

Web | Images | Video | Local | Shopping | more

Search: canon

1 - 20 of about 345,000,000 for canon (About 59k pages)

Also try: [canon digital camera](#), [canon printers](#), [canon cameras](#), [canon usa](#), [More...](#)

**Canon Digital Cameras**  
www.Dealtime.com - Shop and save on everything. Compare products, prices & stores.

**Canon Global (NYSE: CAJ)**  
Global manufacturer of copy machines, fax machines, cameras, computer peripherals, and optical products.  
www.canon.com - 10k - Cached

**Canon USA**  
Manufacturer of professional and consumer imaging equipment and information systems including copiers, printers, image filing systems, cameras and lenses, and more.  
www.usa.canon.com - 59k - Cached

**Canon Cameras**  
Canon offers a range of digital, compact film, and SLR cameras.  
usa.canon.com/consumer/controller/fact/ProductCatIndex1Act&... - 47k - Cached

**Canon Camera Museum**  
Showcasing camera history, technology, and design.  
www.canon.com/camera-museum - 21k - Cached

**Canon UK - Home**  
Specializing in imaging products and solutions for the digital home and digital office environments.  
www.canon.co.uk - 38k - Cached

**Canons - Cheap Prices**  
We Have 800+ Digital Cameras, Canons on Sale, Read Reviews  
www.NextTag.com

**123inkjets.com Coupons & Specials**  
Save Up To 85% on Ink and Toner. Free Shipping Available.  
Give-Me-Ink-Now.com

**Canon Camera Sale**  
Save on Canon Cameras, PowerShot, Rebel & More. Also Camcorders.  
www.Callibex.com/Canon

**Canon IS**  
Canon IS on SALE Save on Digital Cameras.  
best-top-offers.com

[See your message here...](#)

## How do we order Ads ?

- Bid ordering [initially Goto, Overture,...]
- Revenue ordering [now all]
  - by decreasing  $R_x = \text{bid}_x * \text{CTR}_x$

Does Revenue Ordering maximize revenues?

NO - People react to the ordering,  
by changing their bid behavior !!

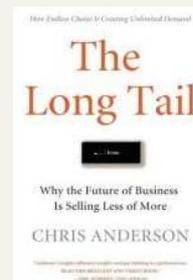
Competitive Market: many advertisers, users  
and the central service [Mechanism Design]

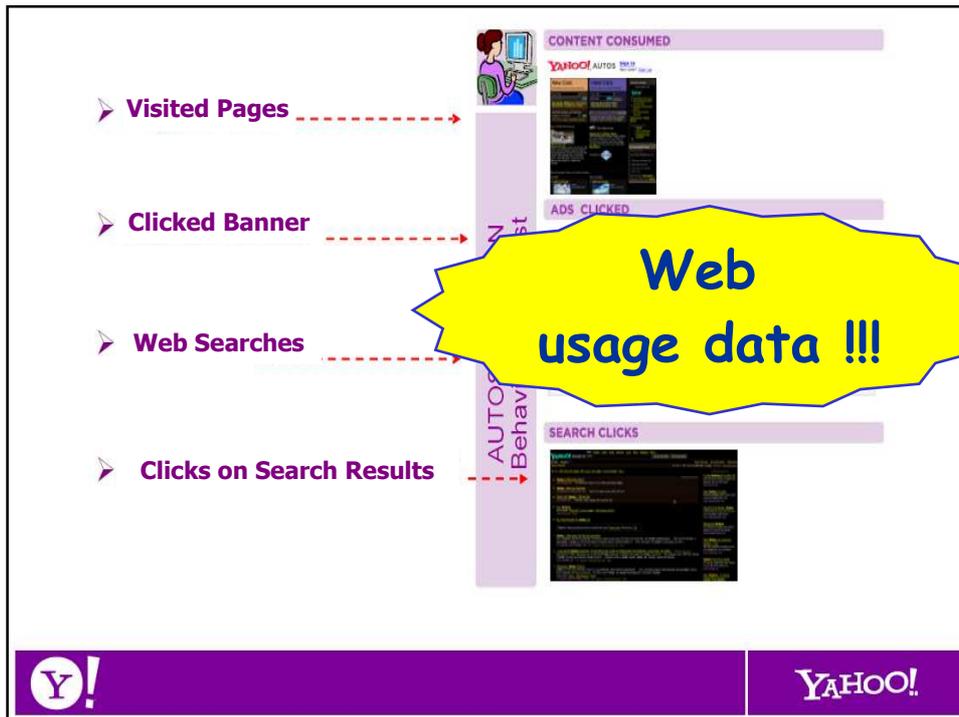
## The new scenario

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

Incentives for specialized niche players

The biggest money is in  
the smallest sales !!





YAHOO! SEARCH [Web](#) | [Images](#) | [Video](#) | [Local](#) | [Shopping](#) | [more...](#)

fc barcelona   [Advanced Search](#)

the Web  just pages in English and Italian

Search Results 1 - 20 of about 4,360,000 for fc barcelona - 0.04 sec. [About this page](#)

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](#)

- F.C. Barcelona**  
 Official site of the Barcelona football club with match information, a photo gallery, and player profiles.  
[www.fcbarcelona.com](http://www.fcbarcelona.com) - [Cached](#)
- 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 Guttmann. ...  
 Quick Links: [History](#) - [Early years \(1899-1908\)](#) - [With Gampel's seal \(1908-1923\)](#)  
[en.wikipedia.org/wiki/FC\\_Barcelona](http://en.wikipedia.org/wiki/FC_Barcelona) - 254k - [Cached](#)
- YouTube - FC BARCELONA**  
 Compilation includes legends like Johan Cruyff, Rivaldo, Romario, ... Gracias FC Barcelona por poner la camiseta sevillista de Antonio Puertal Barça 4 ever. ...  
[www.youtube.com/watch?v=QJ\\_FW-ehUas](http://www.youtube.com/watch?v=QJ_FW-ehUas) - 119k - [Cached](#)
- 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](http://www.fcbarcelona.com/web/english) - 41k - [Cached](#)
- 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](http://commons.wikimedia.org/wiki/Category:FC_Barcelona) - 45k - [Cached](#)
- YouTube - Broadcast Yourself**  
 FC Barcelona. Suscríbete. fcbarcelona. Joined: February 06, 2006. Last Login: 2 days ago ... La página oficial del Fútbol Club Barcelona en YouTube. ...  
[www.youtube.com/fcbarcelona](http://www.youtube.com/fcbarcelona) - 63k - [Cached](#)

**SPONSOR RESULTS**

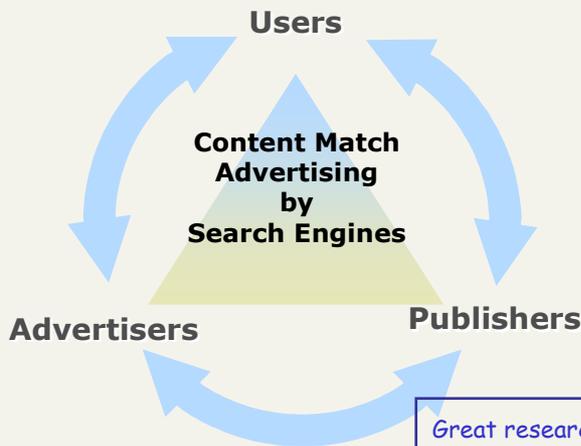
**200 Hotels in Barcelona - Spain**  
 Online hotels in Spain with reviews. Good availability & great rates.  
[www.booking.com](http://www.booking.com)

**Barcelona Hotels**  
 Book your Barcelona Hotel on a Map! Discount online reservations.  
[TopBooker.com](http://TopBooker.com)

[See your message here...](#)

# Ad Industry

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

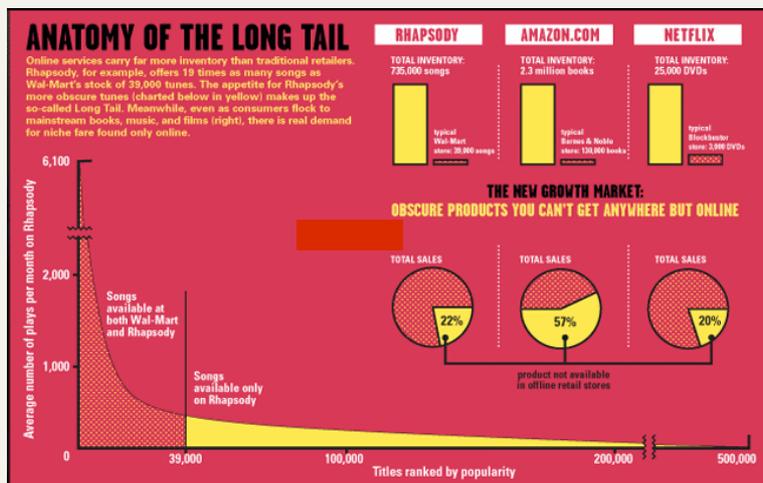


Great research opportunities: we can build the technology to solve the business problem, and we can change the business problem to make it solvable.

# The long tail

## The Long Tail

Why the Future of Business Is Selling Less of More  
CHRIS ANDERSON



(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  $\uparrow$  and  $\downarrow$  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

SEARCH

Books

GO

Computational Molecular Biology: An Algorithmic Approach (Computational Molecular Biology) by Pavel A. Pevzner

List Price: \$47.00  
Price: ~~\$47.00~~ & This item ships for FREE with Super Saver Shipping. See details.

Availability: Usually ships within 24 hours

Used & new from \$42.12

Edition: Hardcover

See more product details

READY TO BUY?

Add to Shopping Cart

or Sign in to turn on 1-Click ordering.

MORE BUYING CHOICES

Used & new from \$42.12

Have one to sell? Sell yours here

Add to Wish List

Add to Wedding Registry

Don't have one? We'll set one up for you.

RECENTLY VIEWED ITEMS

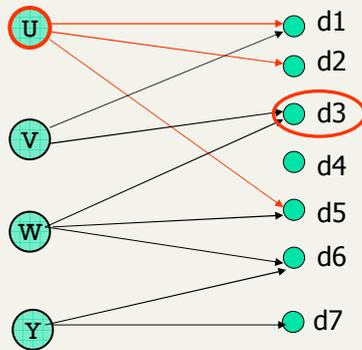
Great Buy

Buy this book with *Algorithms on Strings, Trees, and Sequences* by Dan Gusfield (Author) today!  
Buy Together Today: \$122.00  
Buy both now!

Customers who bought this book also bought:

- Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* by Richard Durbin (Author), et al (Paperback)
- Introduction to Computational Biology: Maps, Sequences and Genomes* by Michael S. Waterman (Paperback)
- Bioinformatics: Sequence and Genome Analysis* by David W. Mount (Hardcover)

# Main idea



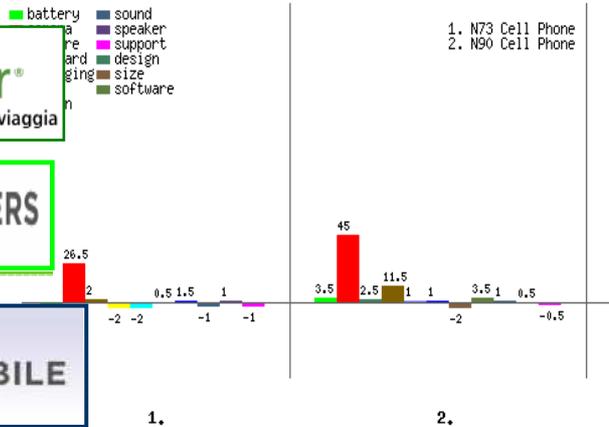
What do we suggest to U ?

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

# An interesting issue



Graphic comparison:



# Web Algorithmics

(The power of Algorithms)

My main interest...

## Solid state hard drives



Traditional hard disk drive

512Bytes → 4 Kb



Solid state hard drive

16Kb → 512 Kb

Fast random reads (60x)  
Slow throughput

## Some tests on PC

[Ajwani et al, 2008]

Experiment	Time/ele SSD	Time/ele HD
Generating & writing $2^{30}$ random doubles	0.4 $\mu$ sec	0.08 $\mu$ sec
Scanning (Read)	0.30 $\mu$ sec	0.09 $\mu$ sec
Sorting	1.6 $\mu$ sec	0.44 $\mu$ sec
Random read	0.4 msec	5.5 msec
Binary Search	1.7 msec	12.4 msec

x5

x3

x4

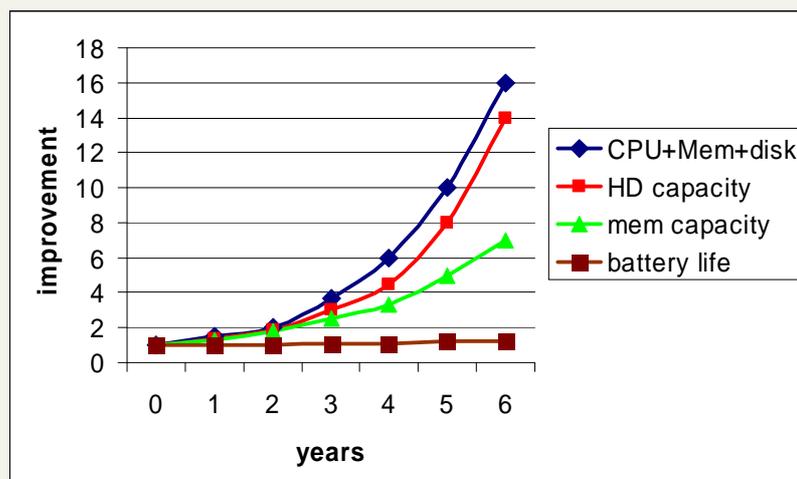
x14

x7 (some locality)

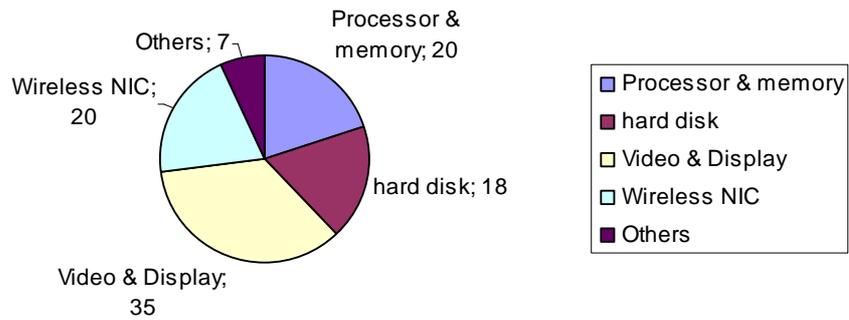
List-ranking a randomly stored list of  $2^{30}$  element list of long integers by naively following pointers takes **minutes** in RAM, but **days** with flash.

Need to distinguish  
between read/write

## Energy vs capacity vs performance



# Energy efficiency



Pizza&Chili Corpus The Italian mirror | The Chilean

Compressed Indexes and their Testbeds

Home | Index Collection | Text Collection | API | Experimental Setup | The Initiative | Additional Material

## The Prologue

The new millennium has seen the born of a new class of *full-text indexes* which are structurally similar to Suffix Trees and Suffix Arrays, in that they support the powerful *substring search* operation, but are *succinct* in space, in that it is close to the empirical entropy of the indexed data. They are therefore called *compressed Suffix Trees* and *compressed Suffix Arrays*, or in general *compressed indexes*.

In the literature we counted more than 20 papers authored by more than 20 different researchers. This interest is motivated by the memory levels and Suffix Arrays.

Don Knuth already closely related tool for the des

**Some figures on a commodity PC:**

- **Count(P)** or **Locate(P)** take few  $\mu\text{sec}/\text{char}$ , in 20÷50% space;
- **Extract** any substring at 1Mb/sec

# 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 !**