World Wide Web (Web, WWW, W3)

WWW

- CERN (the European Organization for Nuclear Research)
  - Tim Berners-Lee
- GUIs
  - Berners-Lee (1990) (HTML, HyperText Transfer Protocol – HTTP, Web Browser), HTTP Web Server
  - Erwise and Viola(1992), Midas (1993) (Initial GUI based Browsers)
- Mosaic (1993)
  - National Center for Supercomputing Applications (NCSA)
  - a hypertext GUI for the X-window system
  - HTML: markup language for rendering hypertext
  - HTTP: hypertext transport protocol for sending HTML and other data over the Internet

http://hal.archives-ouvertes.fr/docs/00/05/44/58/PDF/webgraph.pdf

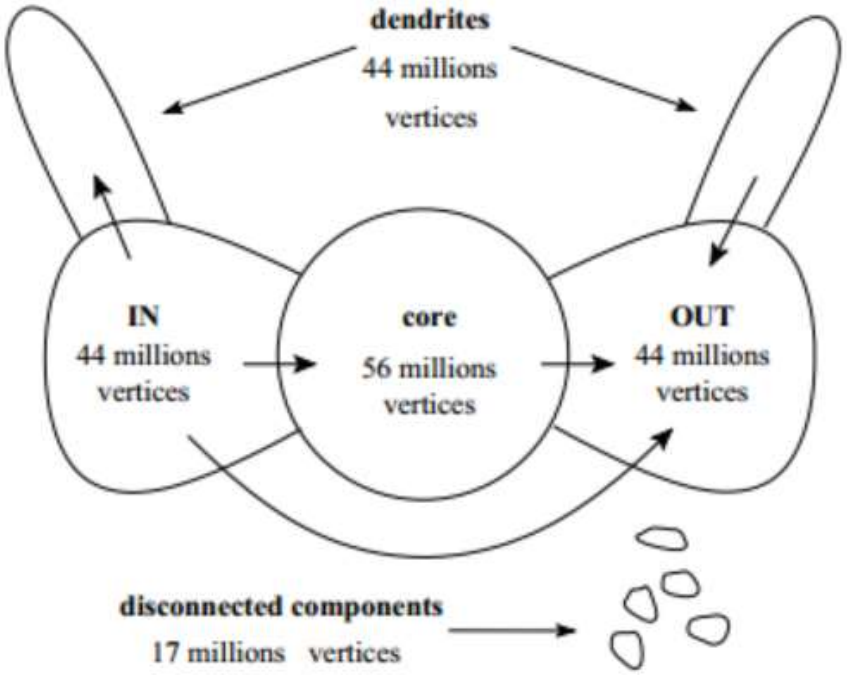
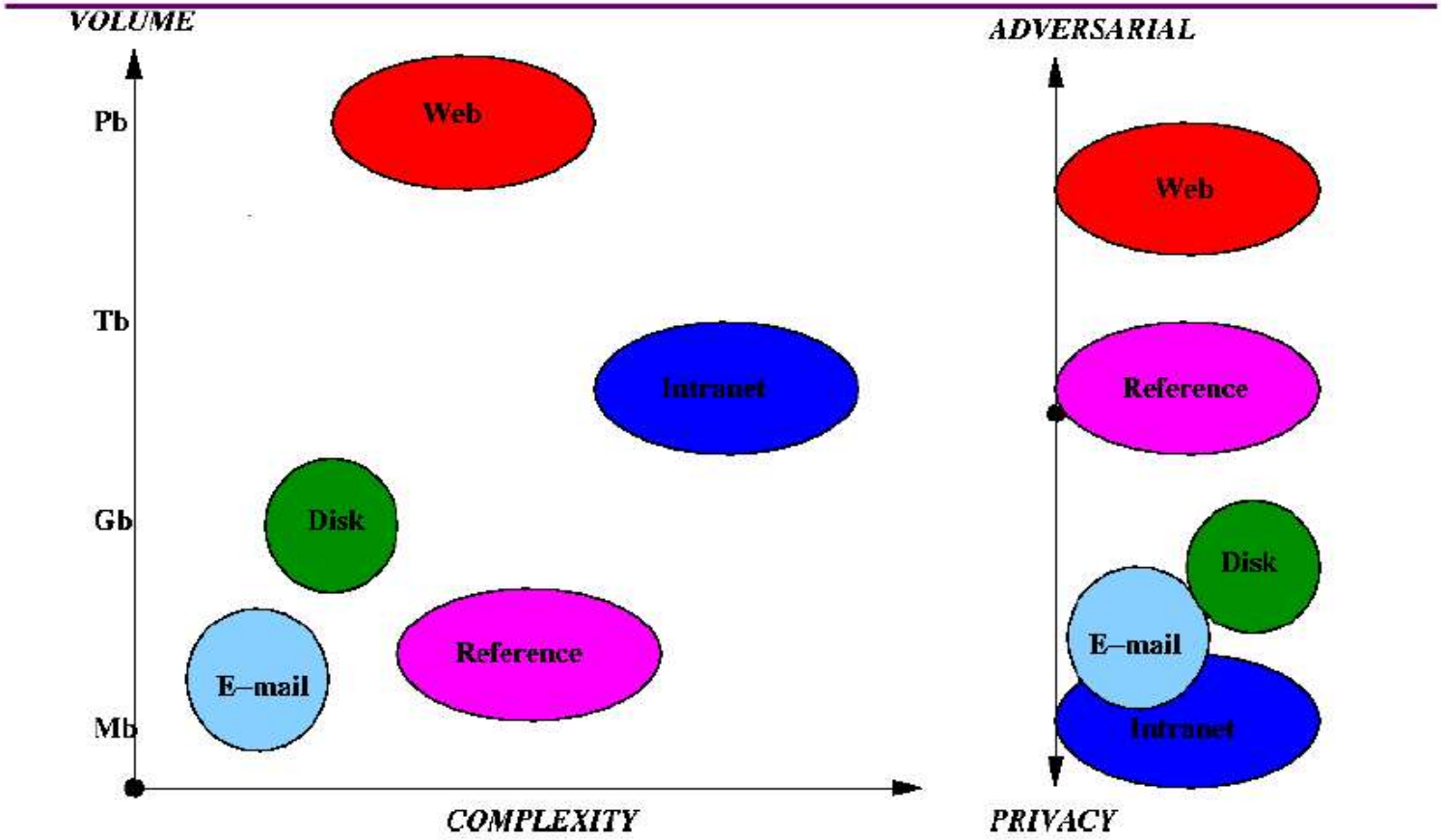


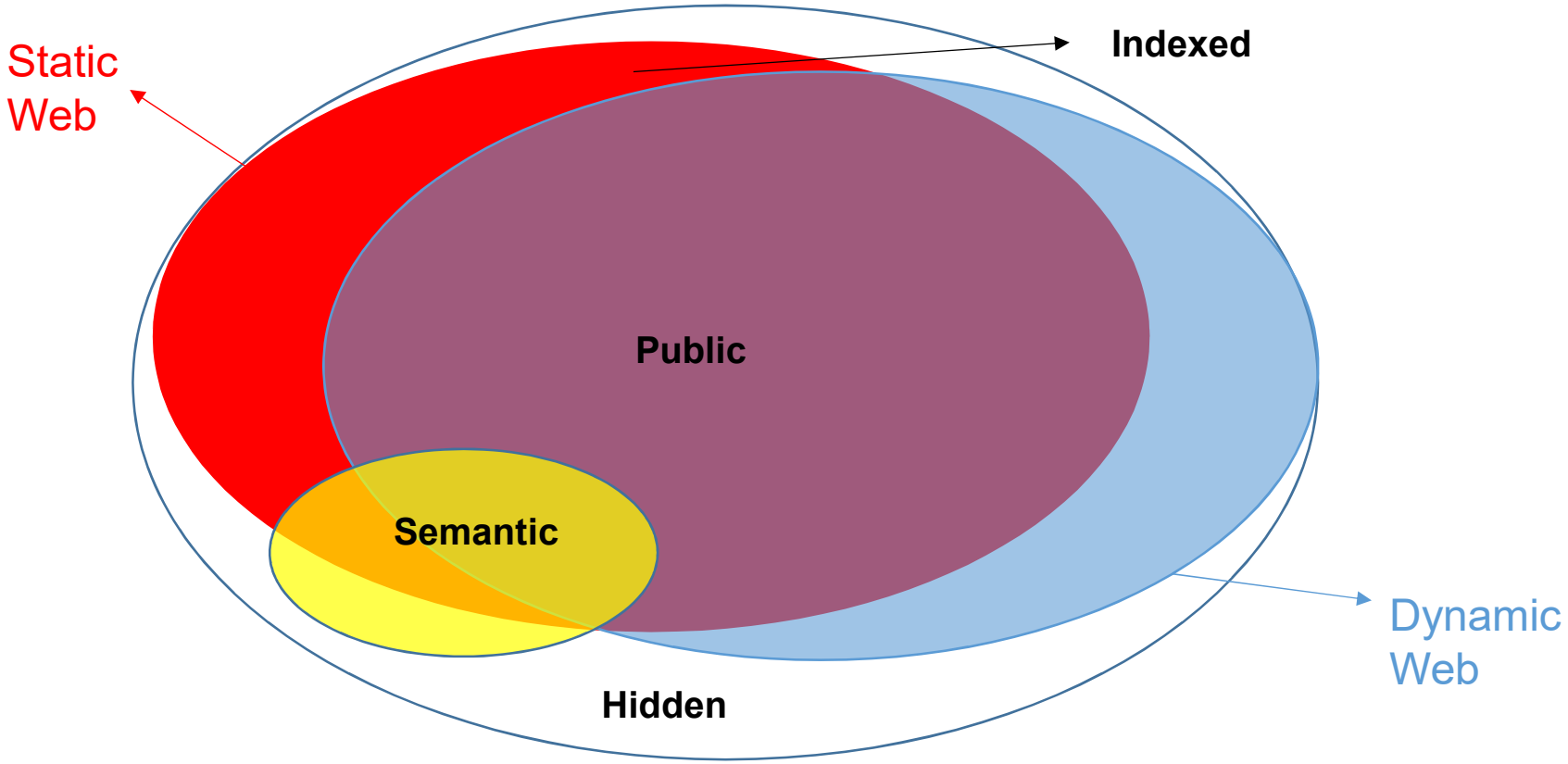
Fig. 3: *The bow-tie macroscopic structure of the Web graph [BKM+00]: the core, the IN component, the OUT component and the dendrites. Each of these parts contains around one quarter of the pages, the disconnected part being reduced to less than 10% of the whole.*

Different Views on Data



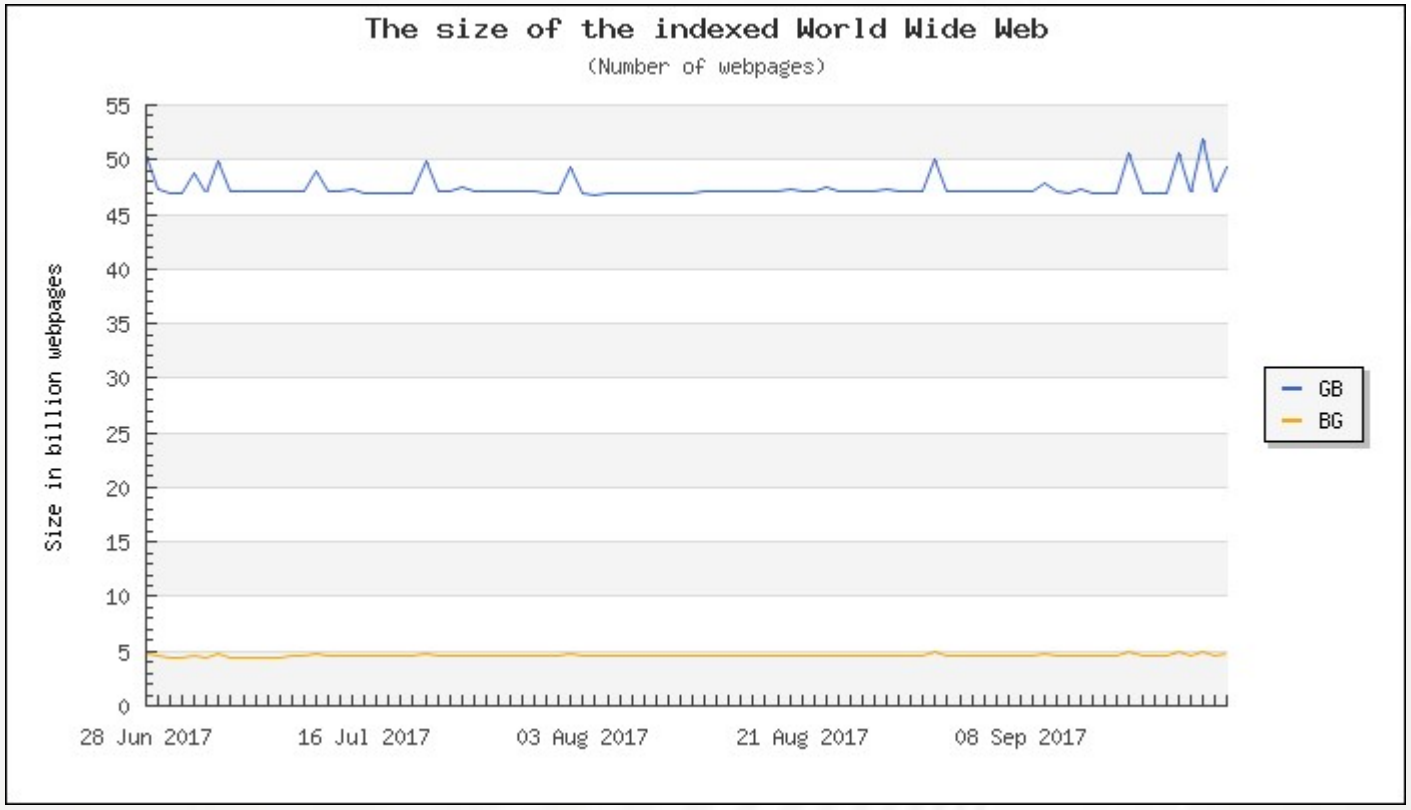
http://ijcai-11.iiia.csic.es/files/proceedings/T16-Web-Mining.pdf

The Different Facets of the Web



http://ijcai-11.iiia.csic.es/files/proceedings/T16-Web-Mining.pdf

How big is the Web?



Web Ne Kadar Büyük?

www.worldwidewebsize.com 26 Eylül 2017

What for?

- The Web as an object
- User-driven Web design
- Improving Web applications
- Social mining
- .....

http://ijcai-11.iiia.csic.es/files/proceedings/T16-Web-Mining.pdf

The Wisdom of Crowds

- James Surowiecki, a *New Yorker* columnist, published this book in 2004
  - "Under the right circumstances, groups are remarkably intelligent"
- Importance of diversity, independence and decentralization

Aggregating data

"large groups of people are smarter than an elite few, no matter how brilliant—they are better at solving problems, fostering innovation, coming to wise decisions, even predicting the future".

http://ijcai-11.iiia.csic.es/files/proceedings/T16-Web-Mining.pdf

http://flickr.com/photos/tags/jaguar/clusters/

Home | Sign Up

Sign In | Help

Photos: Explore Flickr · Learn More

flickr BETA

Tags / jaguar / clusters

jaguar SEARCH

(Or, try an advanced search.)

car, **cars**, **auto**, etype, automobile, classic, vintage, autoshow, red, show

See more in this cluster...

zoo, **animal**, **cat**, animals, bigcat, seattle, woodlandparkzoo, sleep, edinburgh, caged

See more in this cluster...

guitar, **fender**

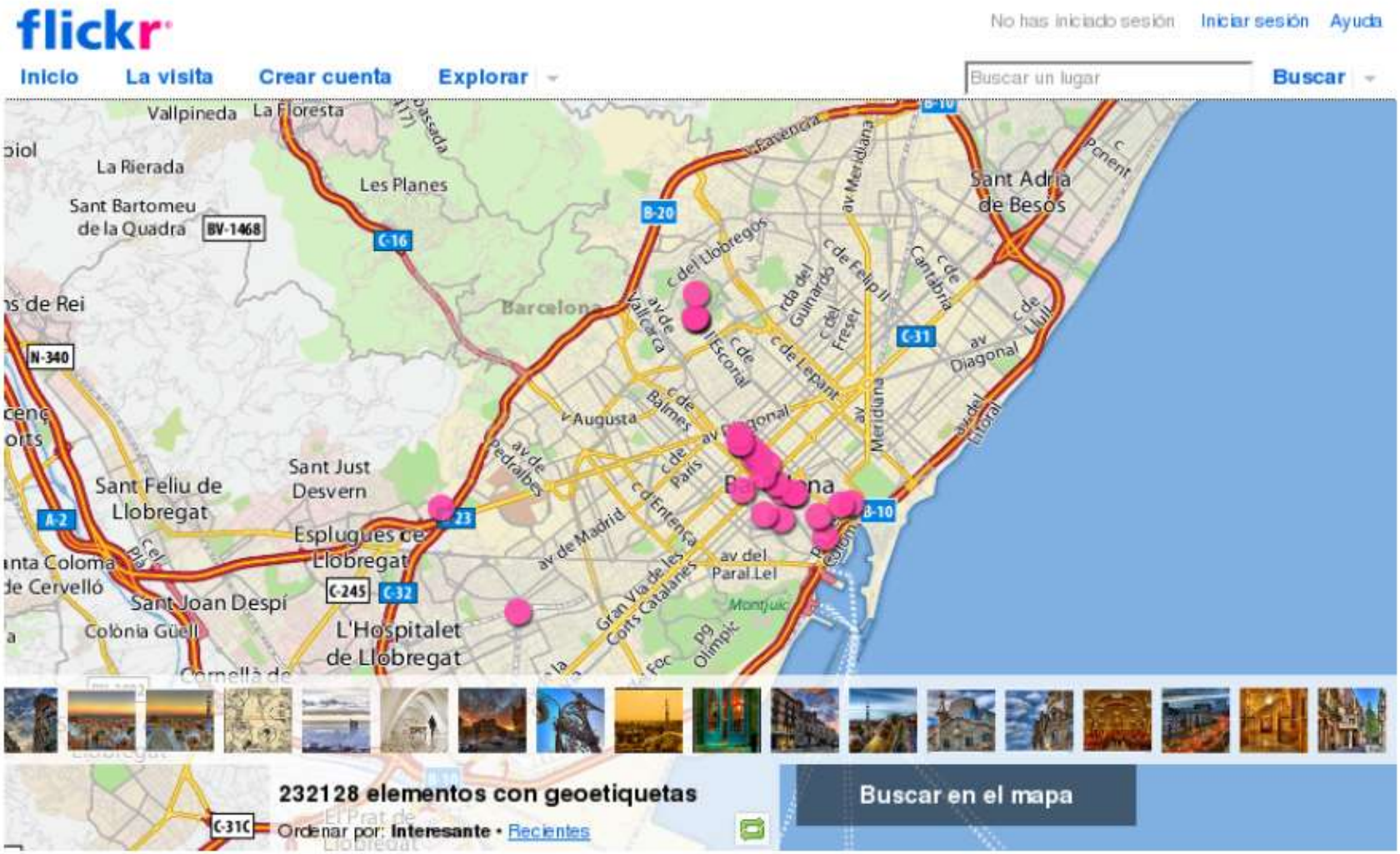
See more in this cluster...

aircraft, **raf**

See more in this cluster...

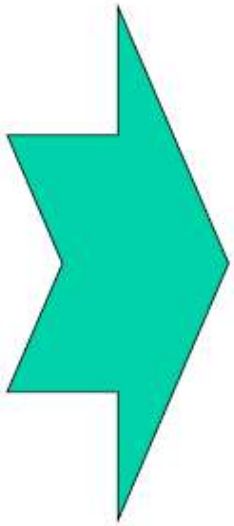
These are the *most recent* photos tagged with **jaguar**. See more...

Flickr: Geo-tagged pictures



The Wisdom of Crowds

- Popularity
- Diversity
- Quality
- Coverage

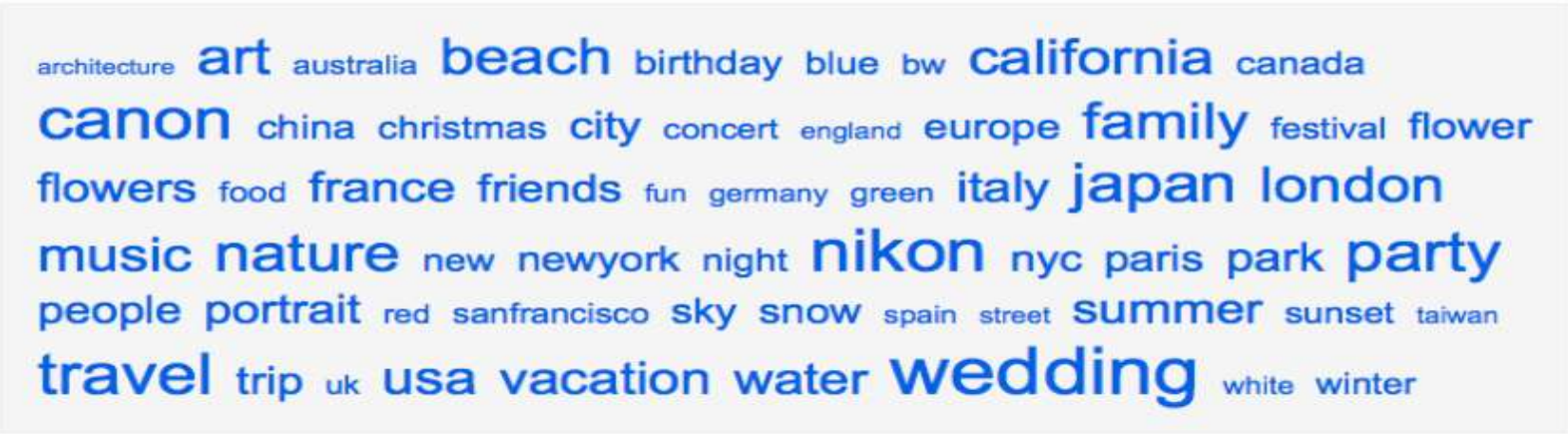


Long tail

http://ijcai-11.iiia.csic.es/files/proceedings/T16-Web-Mining.pdf

The Long Tail

Explore Flickr through tags



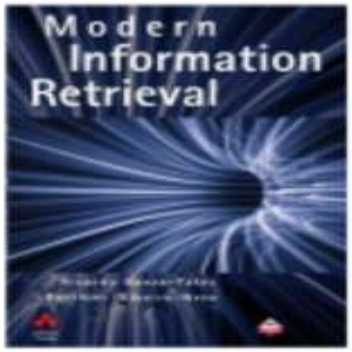
http://ijcai-11.iiia.csic.es/files/proceedings/T16-Web-Mining.pdf

The Wisdom of Crowds

- Crucial for Search Ranking
- Text: Web Writers & Editors
  - –not only for the Web!
- Links: Web Publishers
- Tags: Web Taggers
- Queries: All Web Users!
  - –Queries and actions (or no action!)

http://ijcai-11.iiia.csic.es/files/proceedings/T16-Web-Mining.pdf

What is Information Retrieval (IR)?



IR: Part of computer science which studies the **retrieval of information (not data)** from a collection of **written documents.** The retrieved documents aim at satisfying a **user information need** usually expressed in **natural language.**

- **Documents, unstructured, text, large**
- **Information need**
- **Store, search, find**
- **The World Wide Web?**
- **Relational databases?**

http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws

DIKW

- **Data:** Raw web pages
- **Information:** Result of query
- **Knowledge:** Result of processing query result by user
- **Wisdom:** Synthesis of many such actions by a set of users

- One possible classification of steps in process

Information Retrieval vs. Databases

Information retrieval	Data retrieval
Retrieve all objects **relevant** to some **information need**	Retrieve all objects satisfying some **clearly defined conditions**
Find all documents about the **topic** "semantic web"!	**SELECT** id **FROM** document **WHERE** title **LIKE** '%semantic web%'
Result **list**	**Well-defined** result **set**


```
[selke@tbdb ~]$ db2 "SELECT id FROM document WHERE title L
IKE '%semantic web%' FETCH FIRST 3 ROWS ONLY"

ID
--------------------
               45489
             9635899
               98556

  3 record(s) selected.
```


http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws

Web Search

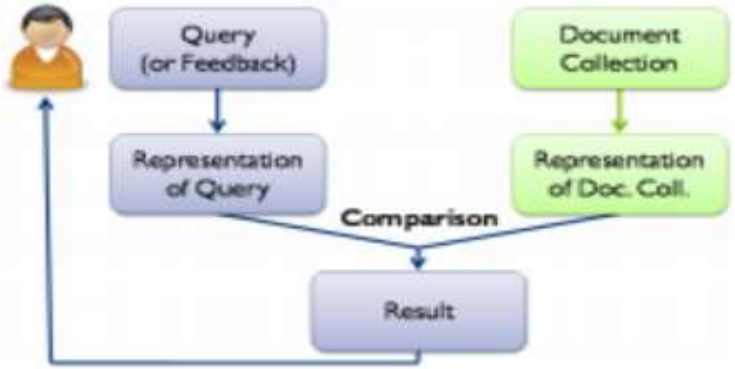
- Very similar to information retrieval
- Main differences:
  - **Links** between Web pages can be exploited
  - **Collecting,** storing, and **updating** documents is more difficult
  - Usually, the **number of users** is very large
  - **Spam** is a problem

5

http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws

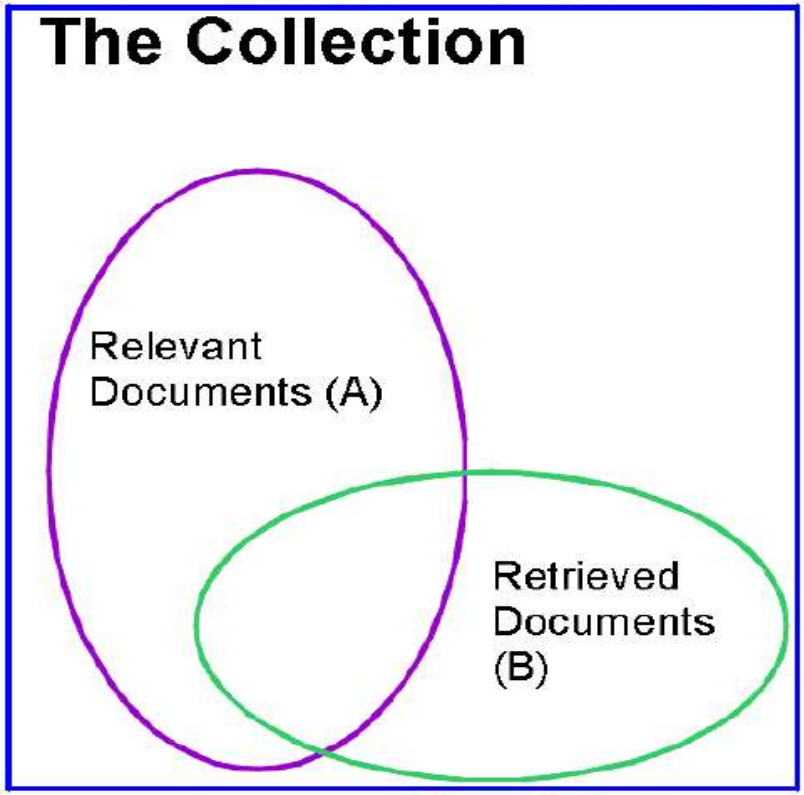
IR Models

- Any IR system is based on an **IR model**
- The model defines …
  - … a **query language,**
  - … an internal **representation of queries,**
  - … an internal **representation of documents,**
  - … a **ranking function** which associates a real number with each query–document pair.
- Optional: A mechanism for **relevance feedback**



http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws

"Classic" Information Retrieval

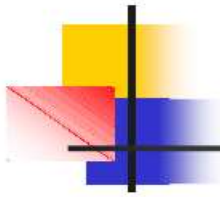


- Given a query, the system retrieves a set B of documents
- Every retrieved document is either relevant or irrelevant to the query

Quality metrics:

- *Recall* : $(A \cap B) / A$
- *Precision* : $(A \cap B) / B$

http://webcourse.cs.technion.ac.il/236375/Winter2013-2014/ho/WCFiles/lec1-intro.pdf



Recall and Precision on the Web

- Relevance of document to queries is not binary – there are many shades of gray
- Broad-topic queries:
  - abundance problem
  - Precision is the dominating factor: users mostly satisfied with a few good results (a few *authoritative* pages)
- Narrow-topic queries:
  - Find a needle in an enormous haystack
  - Recall demands engines cover significant portions of the Web
- Common measure: precision@10
- Nowadays larger emphasis on *diversity*
  - Positive recall for many aspects of the query

http://webcourse.cs.technion.ac.il/236375/Winter2013-2014/ho/WCFiles/lec1-intro.pdf

MA1

New Google Update "Improves Diversity" Of Domains In Results

Sep 14, 2012 at 5:03pm ET by Danny Sullivan

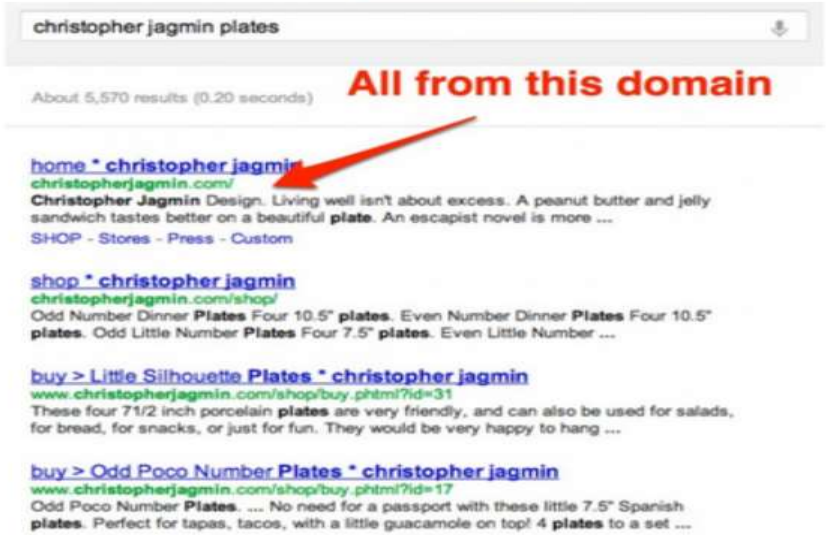
There's been chatter about Google having done some type of algorithm update this week that's having an impact on rankings. Now it's finally confirmed, a new change meant to allow more domains to appear in search results.

The head of Google's spam fighting team Matt Cutts tweeted this about it:

> Just fyi, we rolled out a small algo change this week that improves the diversity of search results in terms of different domains returned.

Lately, more and more people have been noting that Google's search results can sometimes be dominated by pages that all come from the same domain. In other words, do a search, and all the listings seem to come from the same web site.

Here's an example of this from last month, when we wrote about the problem:



Diversity of Search Results

MA1 Mehmet Aktas, 12/15/2013

The Bag of Words Representation

- A very popular **representation of documents** is the **bag of words model**
- Each document is represented by a bag (= multiset) of **terms** from a predefined **vocabulary**
- Standard case:
  - Vocabulary
    = set of all the words occurring in the collection's documents
  - Each document is represented by the words it contains

That's one small step for a man, a giant leap for mankind → { that's, one, small, step, for (2), a (2), man, giant, leap, mankind }

http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws

The Bag of Words Model (2)

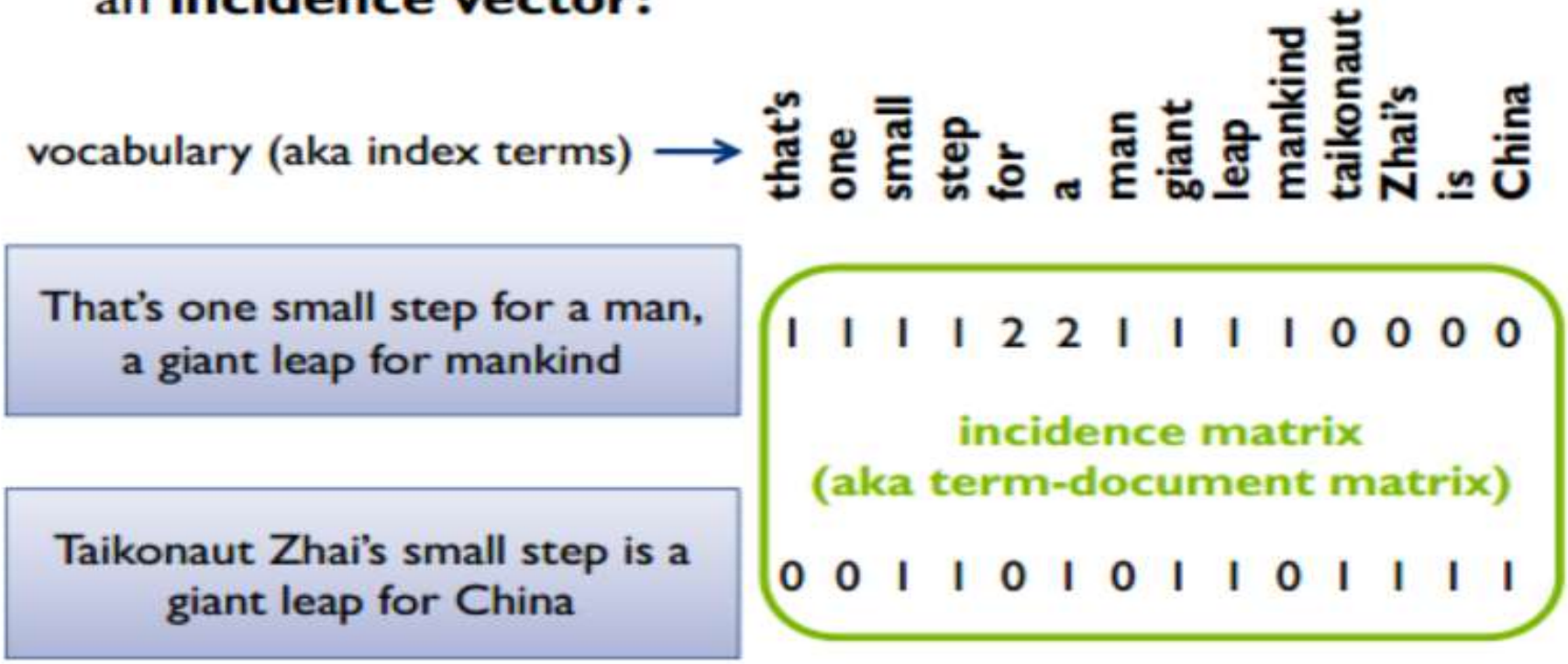
- Cons:
  - Word order gets lost
  - Very different documents could have similar representations
  - Document structure (e.g. headings) and metadata is ignored

- Pros:
  - Simple set-theoretic representation of documents
  - Efficient storage and retrieval of individual terms
  - IR models using the bag of words representation work well!

http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws

The Bag of Words Model (3)

- Any document can be represented by an **incidence vector:**

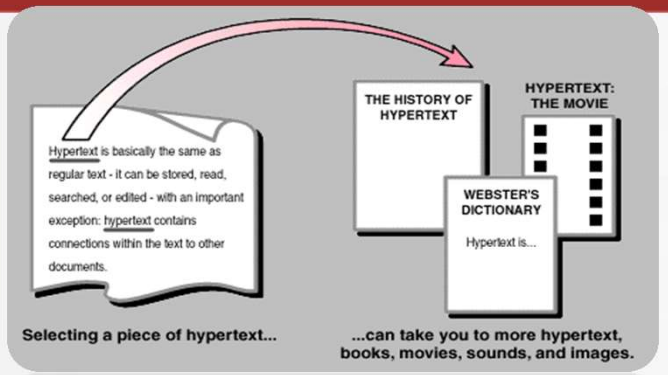


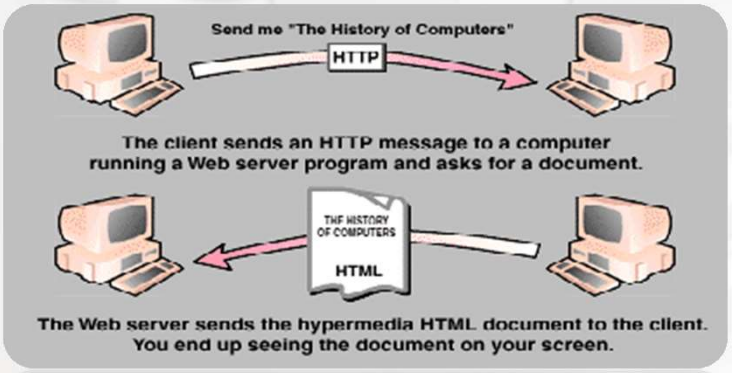
http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws

Web Concepts

- Internet – Web
- Link - Hyperlink
- Hypertext - Hypermedia
- HTTP - HTML
- Anchor Text - URL

Internet vs. Web

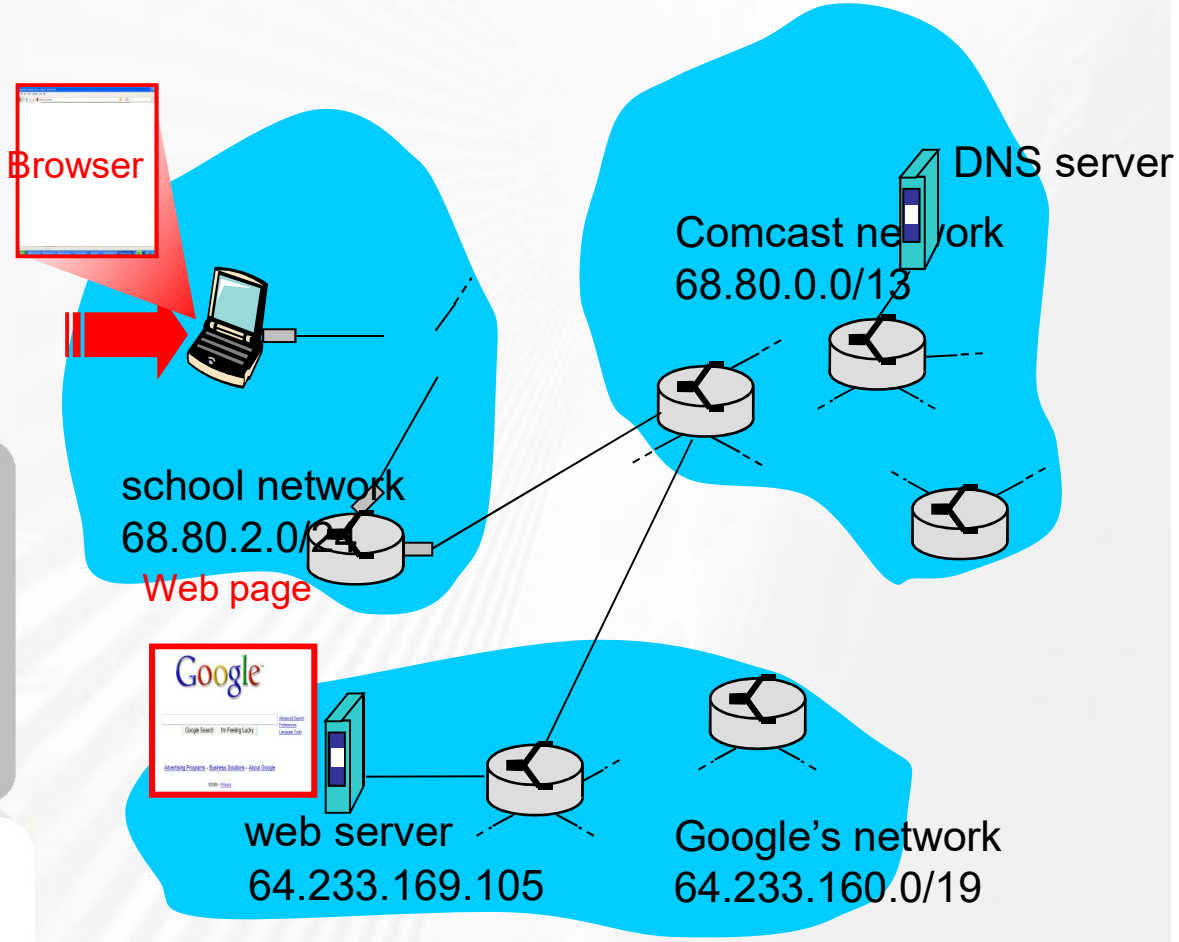




A hyperlink is a link you can click on or activate with the keyboard or other device in order to go somewhere else. The hyperlinks are used on the pages and they navigate from one page to another between same or different Sites/servers.

A URL/Link can be thought of as the "address" of a web page and is sometimes referred to informally as a "web address."

Like if we write "http://google.com" in the address bar, it is a link and if we apply it on the page so that, anyone can navigate from the site to Google, it is called hyperlink.



Anchor text usually gives the user relevant descriptive or contextual information about the content of the link's destination. The anchor text may or may not be related to the actual text of the URL of the link. the English-language Wikipedia's homepage might take this form:

```
<a href="http://en.wikipedia.org/wiki/Main_Page">Wikipedia</a>
```

The anchor text in this example is "Wikipedia"; the unwieldy URL `http://en.wikipedia.org/wiki/Main_Page` displays on the web page as Wikipedia, contributing to clean, easy-to-read text.

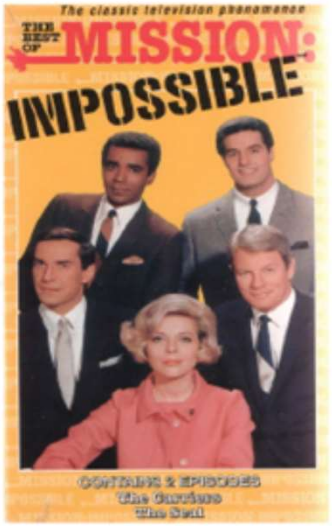
Search Engines

GOOGLE

YANDEX

BING

Mission Impossible?



Search engines:

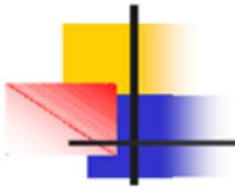
- Crawl and index tens of billions of documents
- Answer hundreds of millions of queries per day
- Devote less than 1 second of processing time per query execution

Users:

- Submit very short queries (averaging about 2.6 terms)
- Expect to receive the most relevant results on the Web
- In a blink of an eye

In terms of 1990 Information Retrieval research - almost unimaginable!

http://webcourse.cs.technion.ac.il/236375/Winter2013-2014/ho/WCFiles/lec1-intro.pdf



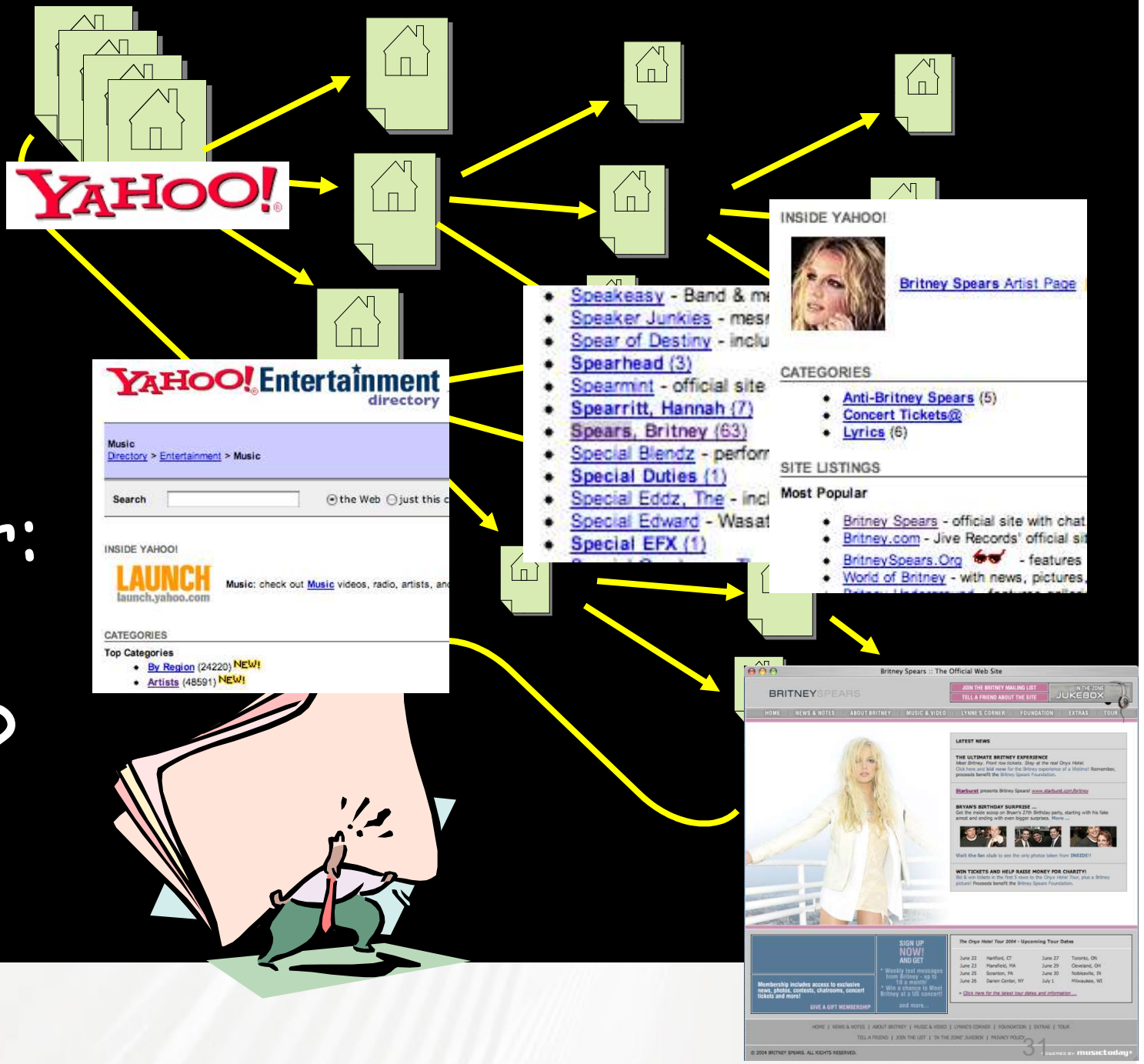
Web Searchers - Observations

- Make ill defined queries
  - Short (2.54 terms average, 80% contain less than 3 words)
  - Use imprecise (and often misspelled) terms
  - Unfamiliar with query syntax (80% queries without operator)
- Wide variance in information needs, expectations, education/knowledge, screen sizes, IP bandwidth, patience
  - Different modalities (mobile, desktop) = different needs and expectations, even with same person
- Specific behavior (desktop and laptop)
  - 85% look over one result screen only (mostly "above the fold")
  - 78% of queries are not modified
- Overall, we as users are investing low cognitive effort per query (formulating and looking at results)

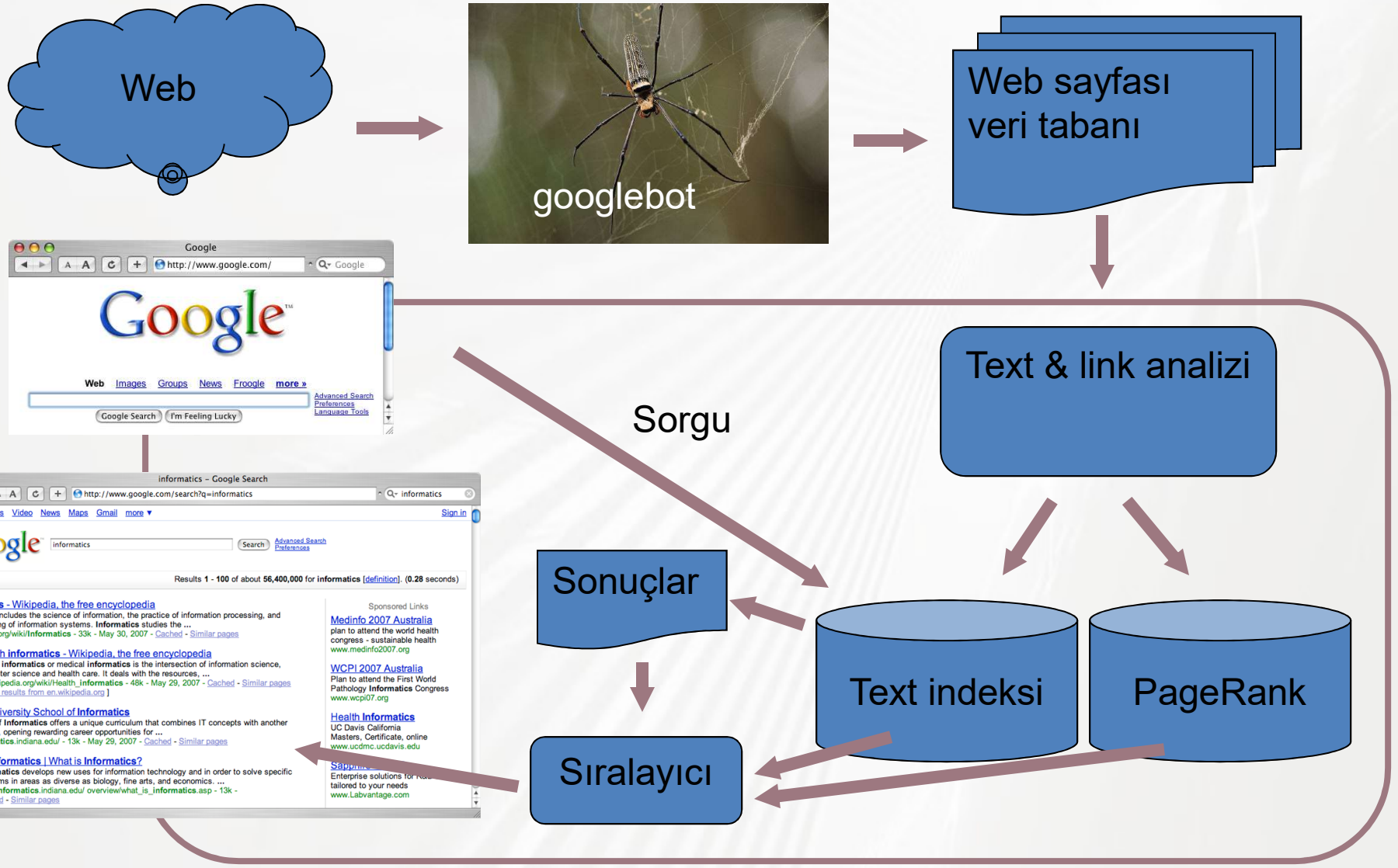
http://webcourse.cs.technion.ac.il/236375/Winter2013-2014/ho/WCFiles/lec1-intro.pdf

Soru: Arama motoru, sonuçlarda listelenen tüm sayfalarda, bu kelimenin olduğunu nasıl biliyor?
Cevap: Tüm bu sayfalar daha önceden indirilip, indeksleniyor.

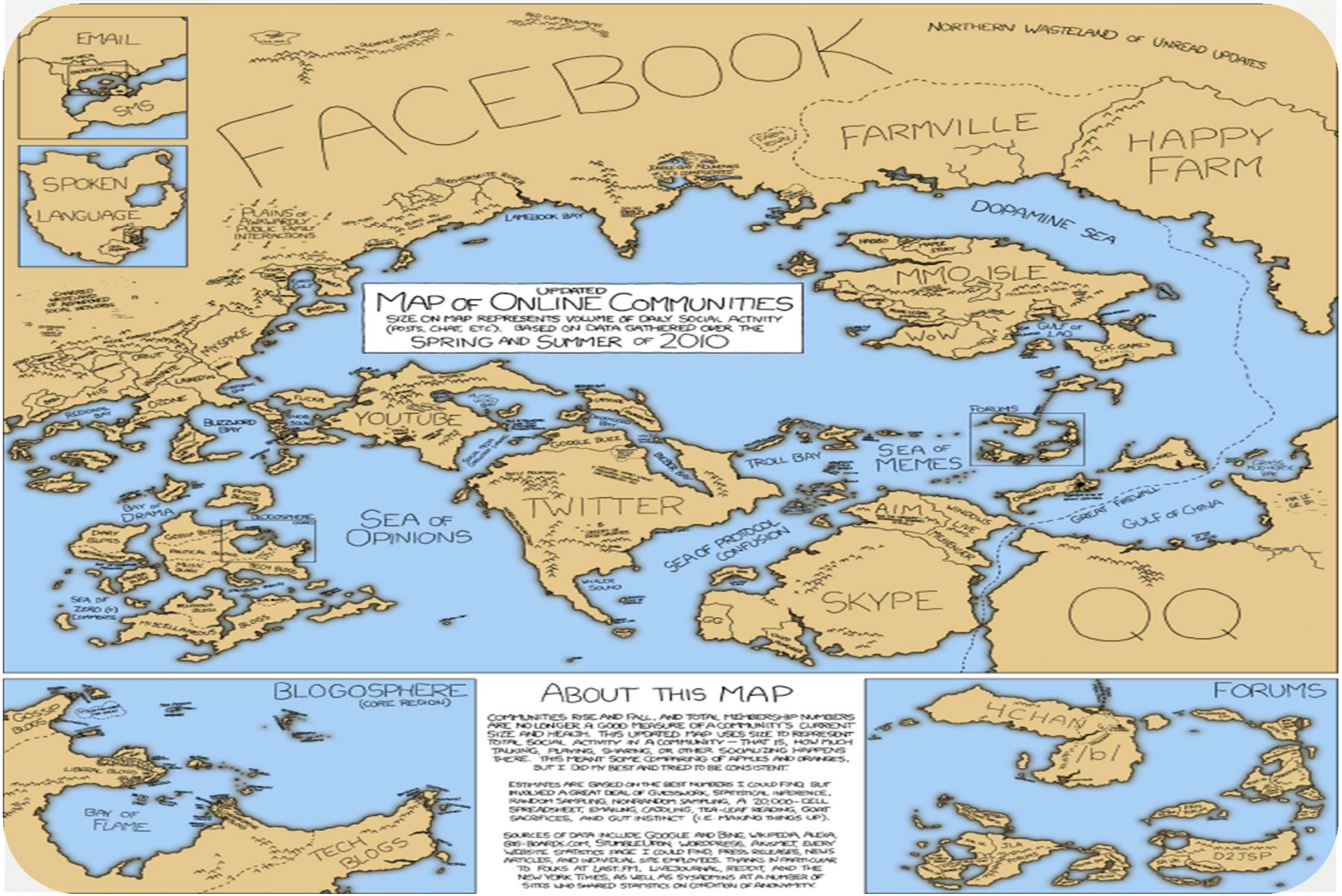
Crawler: Nasıl çalışır?



Crawler


Web
googlebot
Web sayfası veri tabanı
Google
http://www.google.com/
Google
Web Images Groups News Froogle more »
Google Search
I'm Feeling Lucky
Advanced Search
Preferences
Language Tools
Sorgu
Text & link analizi
informatics - Google Search
http://www.google.com/search?q=informatics
Sonuçlar
Text indeksi
PageRank
Sıralayıcı

Online Communities



Why Should I Know about All This?

Gartner

- "80% of business is conducted on **unstructured** information"
- "85% of all data stored is held in an **unstructured** format"
- "7 million **Web pages** are being added every day"

Butler Group
a Datamonitor Company

- "**Unstructured** data doubles every three months"

http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws