Information Retrieval & Data Mining

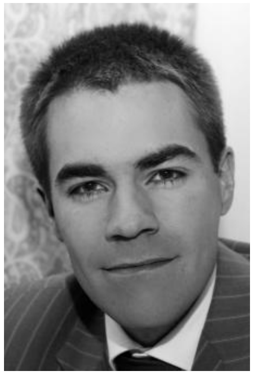
Universität des Saarlandes, Saarbrücken

Winter Semester 2011/12

The Course

Lecturers

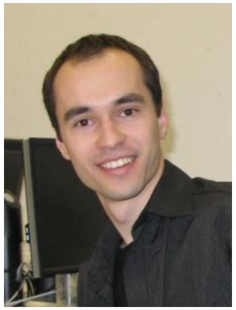
Martin Theobald
martin.theobald@mpi-inf.mpg.de



Pauli Miettinen
pmiettin@mpi-inf.mpg.de

Teaching Assistants

Sarath K. Kondreddi
skondred@mpii.de



Tomasz Tylenda
stylenda@mpii.de

Erdal Kuzey
ekuzey@mpii.de

Mohamed Yahya
myahya@mpii.de

Niket Tandon
ntandon@mpii.de

Faraz Makari
fmakari@mpii.de

D5: Databases & Information Systems Group
Max Planck Institute for Informatics

Organization

- **Lectures:**
  - **Tuesday 14-16** and **Thursday 16-18**

    in **Building E1.3**, **HS-003**
  - Office hours/appointments by e-mail
- **Assignments/tutoring groups**
  - **Friday 12-14, R023, E1.4** (MPI-INF building) *changed from 14-16

    Friday 14-16, SR107, E1.3 (University building)

    Friday 14-16, R023, E1.4 (MPI-INF building) *changed from 16-18

    Friday 16-18, SR016, E1.3 (University building)

  Assignments given out in Thursday lecture, to be solved until next Thursday
  - First assignment sheet given out on **Thursday, Oct 20**
  - First meetings of tutoring groups on **Friday, Oct 28**

Requirements for Obtaining 9 Credit Points

- **Pass 2 out of 3 written tests**

  Tentative dates: **Thu, Nov 17**; **Thu, Dec 22**; **Thu, Jan 26**

  (45-60 min each)

- **Pass the final written exam**

  Tentative date: **Tue, Feb 21** (120-180 min)

- Must **present solutions to 3 assignments**, more possible

  (**You must return your assignment sheet and have a correct solution in order to present in the exercise groups.**)

  - **1 bonus point** possible in tutoring groups
  - **Up to 3 bonus points** possible in tests
  - Each bonus point earns one mark in letter grade (0.3 in numerical grade)

Register for Tutoring Groups

http://www.mpi-inf.mpg.de/departments/d5/teaching/ws11_12/irdm/

- Register for one of the tutoring groups **until Oct. 27**
- Check back frequently for updates & announcements

Agenda

I. Introduction

II. Basics from probability theory & statistics

III. Ranking principles

IV. Link analysis

V. Indexing & searching

VI. Information extraction

Information Retrieval

VII. Frequent item-sets & association rules

VIII. Unsupervised clustering

IX. (Semi-)supervised classification

X. Advanced topics in data mining

Data Mining

XI. Wrap-up & summary

Literature (I)

- **Information Retrieval**

  - Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze.
    Introduction to Information Retrieval
    Cambridge University Press, 2008.
    Website: http://nlp.stanford.edu/IR-book/

  - R. Baeza-Yates, R. Ribeiro-Neto.
    Modern Information Retrieval: The concepts and technology behind search.
    Addison-Wesley, 2010.

  - W. Bruce Croft, Donald Metzler, Trevor Strohman.
    Search Engines: Information Retrieval in Practice.
    Addison-Wesley, 2009.
    Website: http://www.pearsonhighered.com/croft1epreview/

Literature (II)

- **Data Mining**
  - Mohammed J. Zaki, Wagner Meira Jr.
    Fundamentals of Data Mining Algorithms
    Manuscript (will be made available during the semester)
  - Pang-Ning Tan, Michael Steinbach, Vipin Kumar.
    Introduction to Data Mining
    Addison-Wesley, 2006.
    Website: http://www-users.cs.umn.edu/%7Ekumar/dmbook/index.php

Literature (III)

- **Background & Further Reading**

  - Jiawei Han, Micheline Kamber, Jian Pei.
    Data Mining - Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2011
    Website: http://www.cs.sfu.ca/~han/dmbook

  - Stefan Büttcher, Charles L. A. Clarke, Gordon V. Cormack.
    Information Retrieval: Implementing and Evaluating Search Engines,
    MIT Press, 2010

  - Christopher M. Bishop.
    Pattern Recognition and Machine Learning, Springer, 2006.

  - Larry Wasserman.
    All of Statistics, Springer, 2004.
    Website: http://www.stat.cmu.edu/~larry/all-of-statistics/

  - Trevor Hastie, Robert Tibshirani, Jerome Friedman.
    The elements of statistical learning, 2nd edition, Springer, 2009.

Quiz Time!

- Please answer the **20 quiz questions** during the rest of the lecture.
- The quiz is completely **anonymous**, but keep your id on the top-right corner. There will be a **prize for the 3 best answer** sheets.

Chapter I: Introduction – IRDM Applications & System Architectures

Information Retrieval & Data Mining

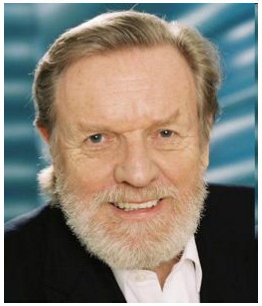
Universität des Saarlandes, Saarbrücken

Winter Semester 2011/12

Chapter I: IRDM Applications and System Architectures

- **1.1 Overview of IRDM Technologies & Applications**
- **1.2 Search Engines – IR in a Nutshell**
  - Deep Web / Hidden Web, Semantic Web, Multimodal Web, Social Web (Web 2.0)
- **1.3 Data Mining in a Nutshell**
  - Real-world DM applications

„We are drowning in information, and starved for knowledge.“
-- John Naisbitt



I.1 Overview of Applications & Technologies

Objective: Satisfy information demand & curiosity of human users – and eliminate the (expensive) bottleneck of human time !

Information Retrieval (IR):

- document content & structure analysis
- indexing, search, relevance ranking
- classification, grouping, segmentation
- interaction with knowledge bases
- annotation, summarization, visualization
- personalized interaction & collaboration

application areas:

- Web & Deep Web search
- digital libraries & enterprise search
- XML & text integration
- multimedia information
- Web 2.0 and social networks
- personalized & collaborative filtering

Data Mining (DM):

- learning predictive models from data
- pattern, rule, trend, outlier detection
- classification, grouping, segmentation
- knowledge discovery in data collections
- information extraction from text & Web
- graph mining (e.g. on Web graph)

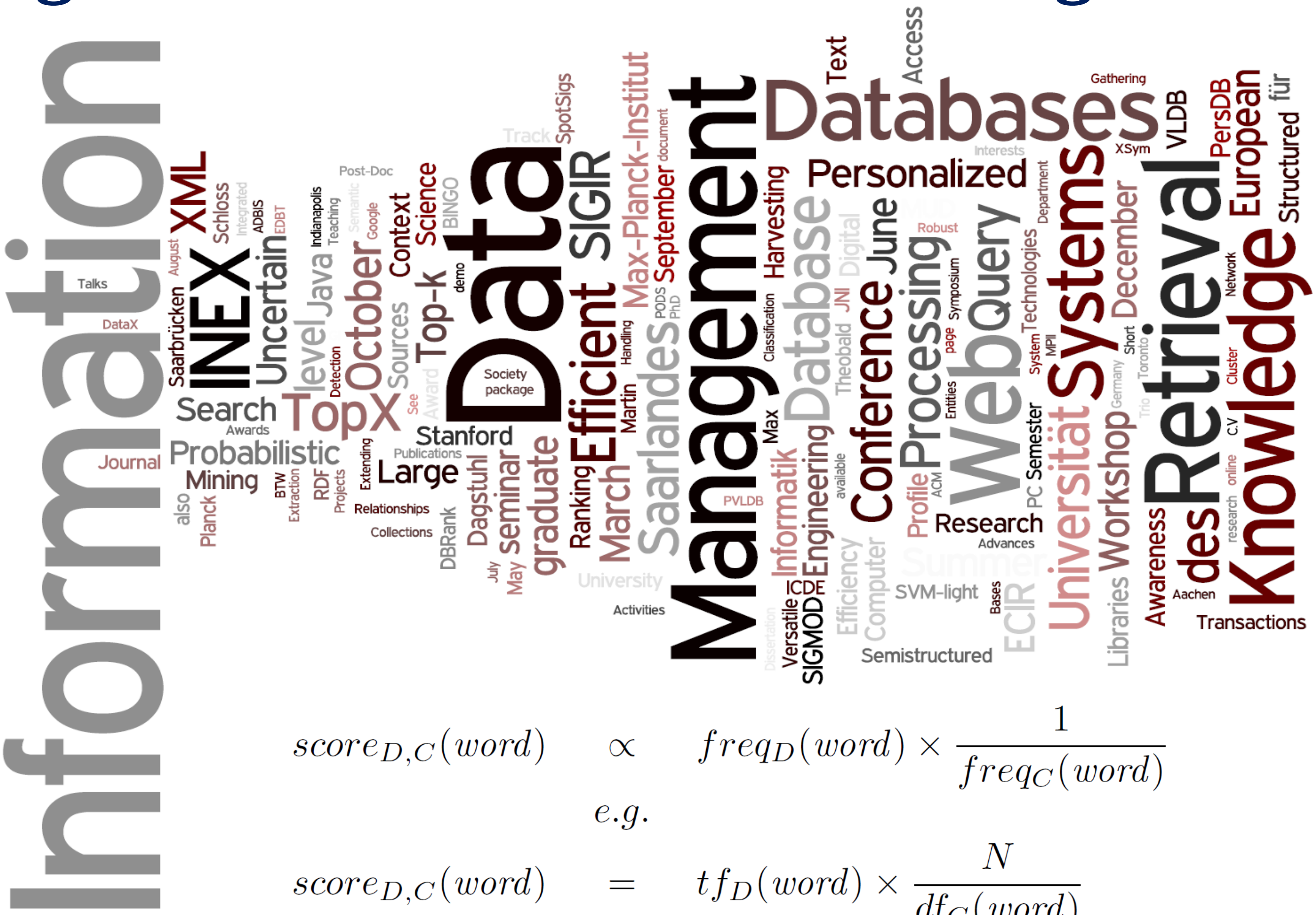
application areas:

- bioinformatics, e.g.: protein folding, medical therapies, gene co-regulation
- business intelligence, e.g.: market baskets, CRM, loan or insurance risks
- scientific observatories, e.g.: astrophysics, Internet traffic (incl. fraud, spam, DoS)
- Web mining & knowledge harvesting

Connected to natural language processing (NLP) and statistical machine learning (ML)

Tag Clouds – Retrieval or Mining?

http://www.wordle.net/



$$score_{D,C}(word) \propto freq_D(word) \times \frac{1}{freq_C(word)}$$

$$e.g.$$

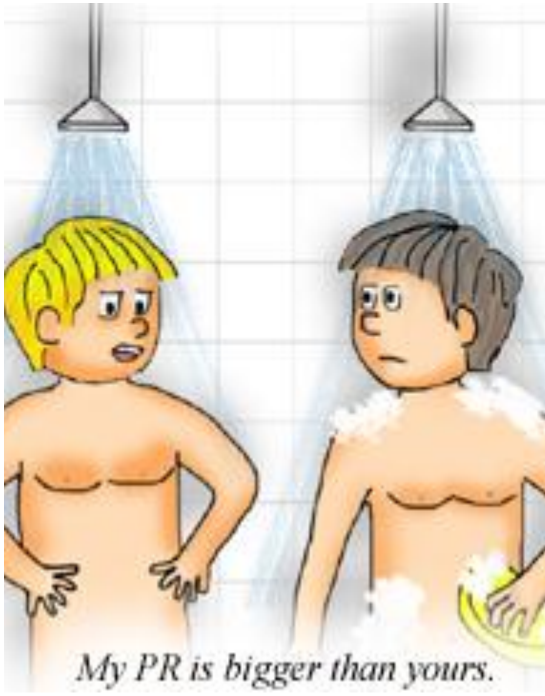
$$score_{D,C}(word) = tf_D(word) \times \frac{N}{df_C(word)}$$

The Google Revolution

Actually, I think I have enough keywords now to consult Google.

© Philipp Lenssen
blog.outer-court.com

© Philipp Lenssen
blog.outer-court.com

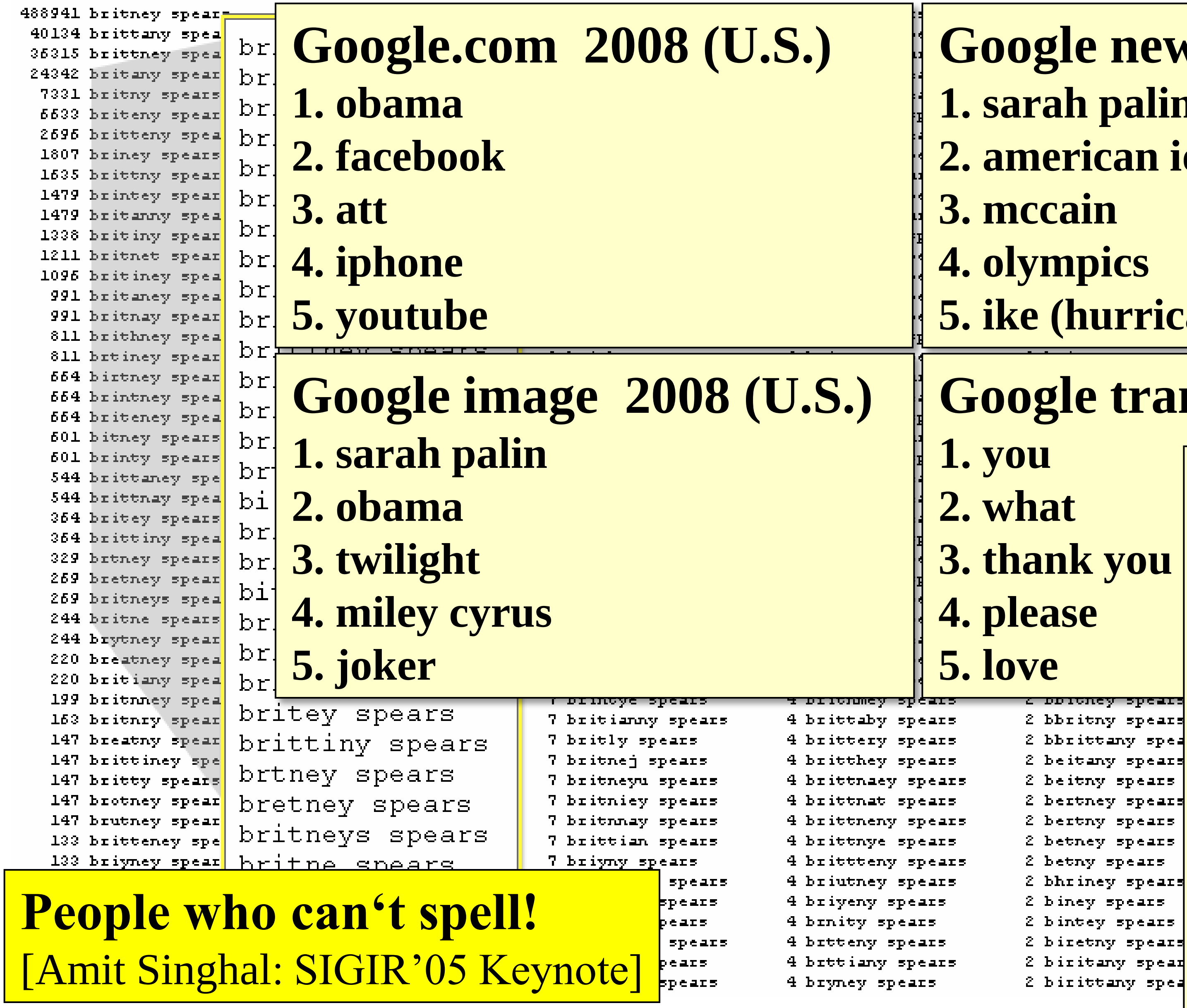


© 2003 Verity Intellectual Properties Pty Ltd

- ★ great for e-shopping, school kids, scientists, doctors, etc.
- ★ high-precision results for simple queries
- ★ superb scalability & throughput (> 20 Bio. docs, > 1000 queries/sec)
- ★ continuously enhanced: GoogleScholar, GoogleEarth, Google+, multilingual for >100 languages, calendar, query auto-completion,…

Search Engine Users

http://www.google.com/press/zeitgeist2010/regions/de.html



```
488941 britney spears
 40134 brittany spea
 36315 brittney spea
 24342 britany spear
  7331 britny spears
  6633 briteny spear
  2696 britteny spea
  1807 briney spears
  1635 brittny spear
  1479 brintey spear
  1479 britanny spea
  1338 britiny spear
  1211 britnet spear
  1096 britiney spea
   991 britaney spea
   991 britnay spear
   811 brithney spea
   811 brtiney spear
   664 birtney spear
   664 brintney spea
   664 briteney spea
   601 bitney spears
   601 brinty spears
   544 brittaney spe
   544 brittnay spea
   364 britey spears
   364 brittiny spea
   329 brtney spears
   269 bretney spear
   269 britneys spea
   244 britne spears
   244 brytney spear
   220 breatney spea
   220 britiany spea
   199 britnney spea
   163 britnry spear
   147 breatny spear
   147 brittiney spe
   147 britty spears
   147 brotney spear
   147 brutney spear
   133 britteney spe
   133 briyney spear
```

```
britey spears
brittiny spears
brtney spears
bretney spears
britneys spears
britne spears
```

```
7 britianny spears
7 britly spears
7 britnej spears
7 britneyn spears
7 britniey spears
7 britnnay spears
7 brittian spears
7 briyny spears
```

```
4 brittaby spears
4 brittery spears
4 britthey spears
4 brittnaey spears
4 brittnat spears
4 brittneny spears
4 brittnye spears
4 britttteny spears
4 briutney spears
4 briyeny spears
4 brnity spears
4 brtteny spears
4 brttiany spears
4 bryney spears
```

```
2 bbritny spears
2 bbrittany spea
2 beitany spears
2 beitny spears
2 bertney spears
2 bertny spears
2 betney spears
2 betny spears
2 bhriney spears
2 biney spears
2 bintey spears
2 biretny spears
2 biritany spear
2 birittany spea
```

Google.com 2008 (U.S.)

1. obama
2. facebook
3. att
4. iphone
5. youtube

Google news 2008 (U.S.)

1. sarah palin
2. american idol
3. mccain
4. olympics
5. ike (hurricane)

Google image 2008 (U.S.)

1. sarah palin
2. obama
3. twilight
4. miley cyrus
5. joker

Google translate 2008 (U.S.)

1. you
2. what
3. thank you
4. please
5. love

Google.de 2008

1. wer kennt wen
2. juegos
3. facebook
4. schüler vz
5. studi vz
6. jappy
7. youtube
8 yasni
9. obama
10. euro 2008

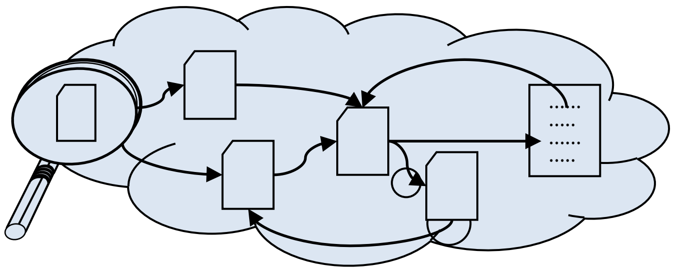
People who can‘t spell!

[Amit Singhal: SIGIR’05 Keynote]

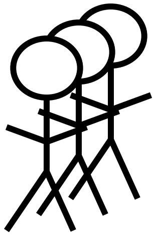
Web Search Patterns [Rose/Levinson: WWW 2004]

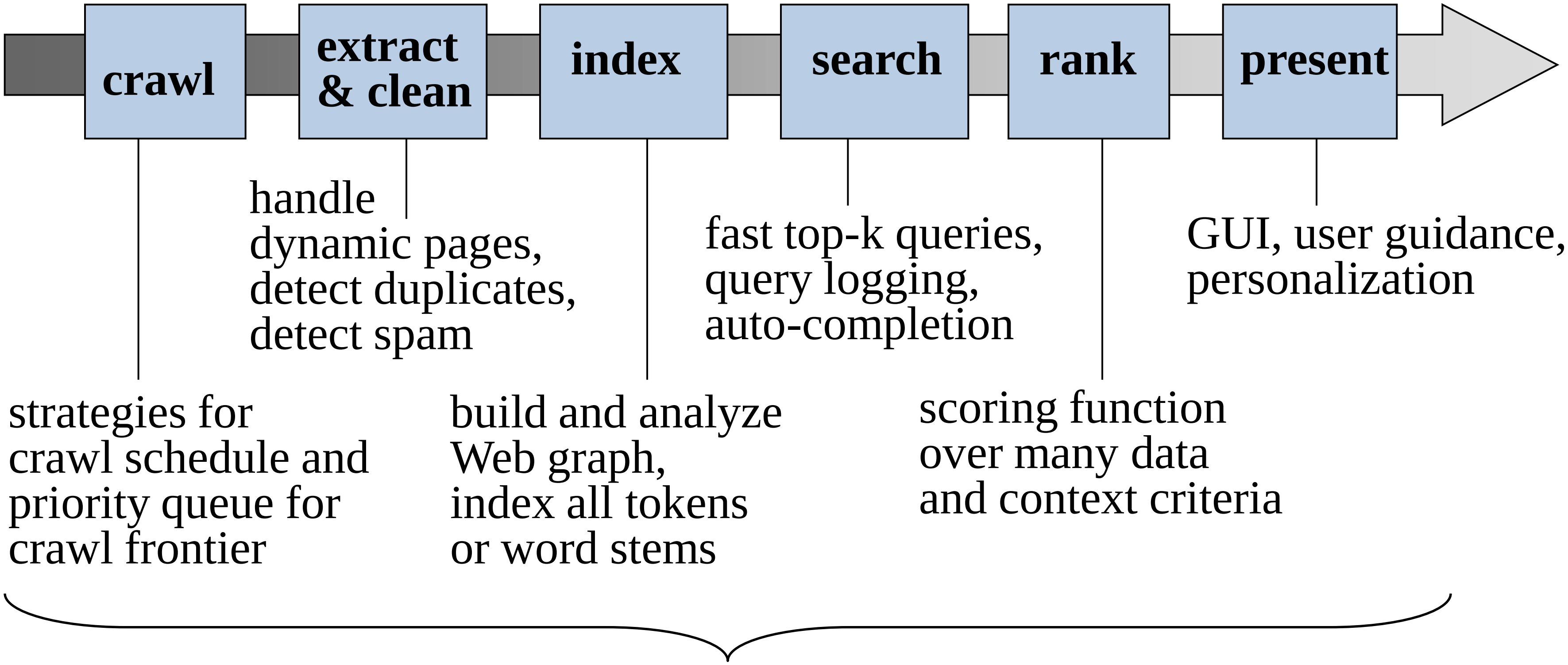
- **navigational:** find specific homepage with unknown URL, e.g. Cirrus Airlines
- **transactional:** find specific resource, e.g. download Lucene source code, Sony Cybershot DSC-W5, Mars surface images, hotel beach south Crete August
- **informational:** learn about topic
  - focused, e.g. Chernoff bounds, soccer world championship qualification
  - unfocused, e.g. undergraduate statistics, dark matter, Internet spam
  - seeking advice, e.g. help losing weight, low-fat food, marathon training tips
  - locating service, e.g. 6M pixel digital camera, taxi service Saarbrücken
  - exhaustive, e.g. Dutch universities, hotel reviews Crete, MP3 players
- **embedded in business workflow** (e.g. CRM, business intelligence) or **personal agent** (in cell phone, MP3 player, or ambient intelligence at home) **with automatically generated queries**
- **natural-language question answering (QA):**
  - **factoids**, e.g. when was Johnny Depp born, where is the Louvre, who is the CEO of Google, what kind of particles are quarks, etc.
  - **list queries**, e.g. in which movies did Johnny Depp play
  - **opinions**, e.g. Barack Obama, should Germany leave Afghanistan, etc.

I.2 Search Engines (IR in a Nutshell)



- Web, intranet, digital libraries, desktop search
- Unstructured/semistructured data





handle dynamic pages, detect duplicates, detect spam

fast top-k queries, query logging, auto-completion

GUI, user guidance, personalization

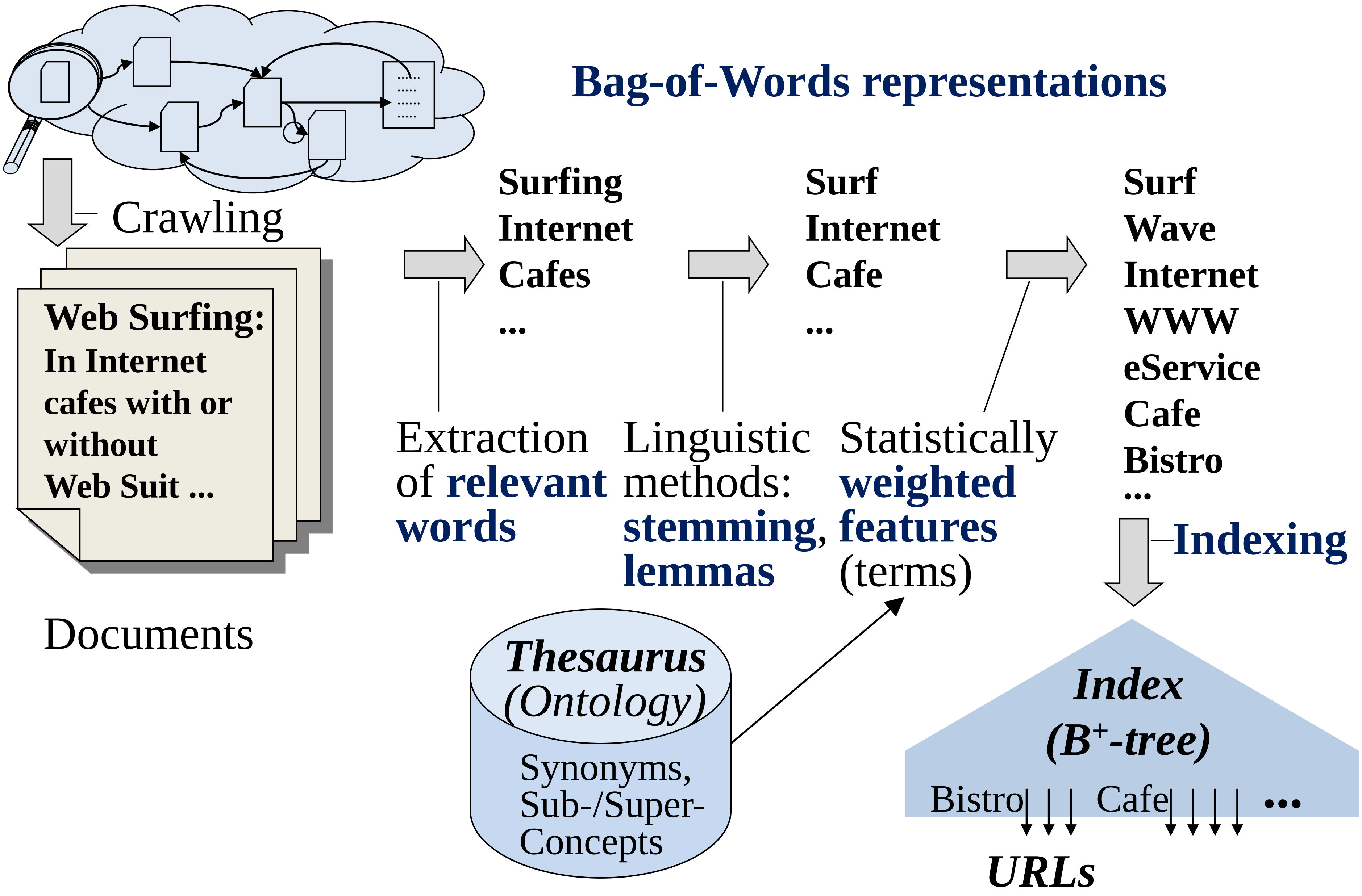
strategies for crawl schedule and priority queue for crawl frontier

build and analyze Web graph, index all tokens or word stems

scoring function over many data and context criteria

Server farms with **10 000's** (2002) – **100,000's** (2010) computers,
distributed/replicated data in high-performance file system (**GFS**,**HDFS**,…),
massive parallelism for query processing (**MapReduce**, **Hadoop**,…)

Content Gathering and Indexing



Bag-of-Words representations

Crawling

Web Surfing:
In Internet cafes with or without Web Suit ...

Documents

Surfing
Internet
Cafes
...

Extraction of **relevant words**

Surf
Internet
Cafe
...

Linguistic methods: **stemming**, **lemmas**

Statistically **weighted features** (terms)

Surf
Wave
Internet
WWW
eService
Cafe
Bistro
...

Indexing

Thesaurus
(Ontology)

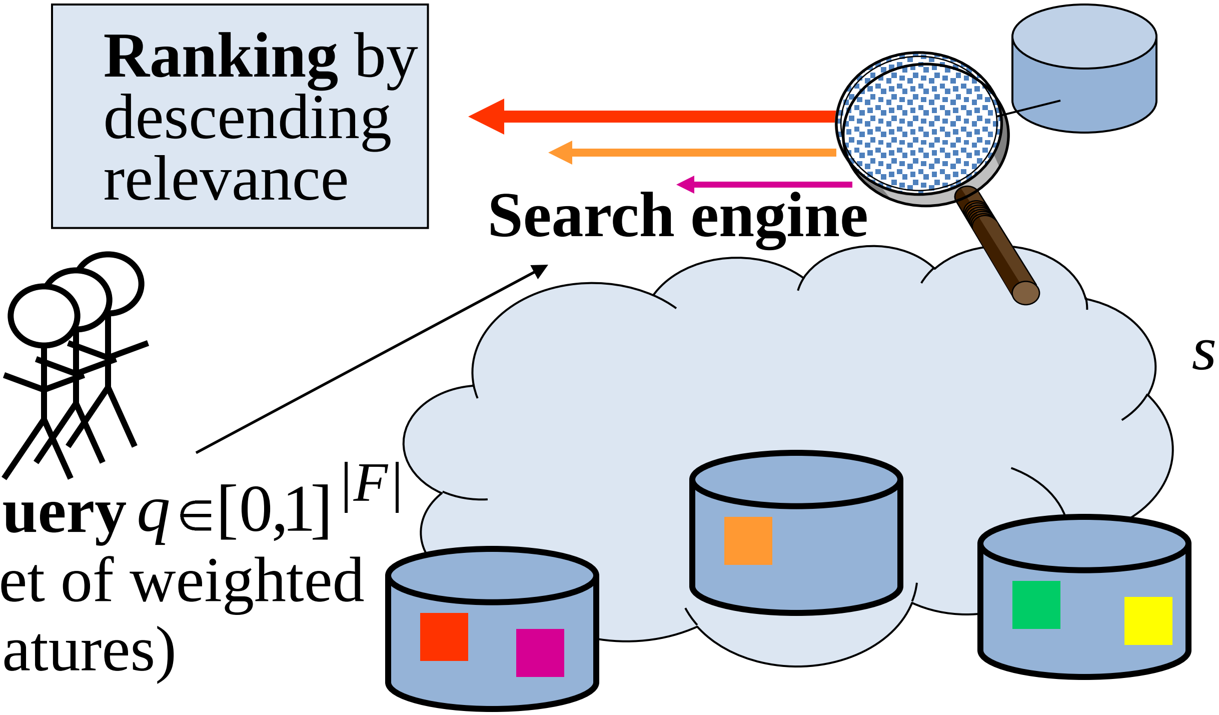
Synonyms,
Sub-/Super-
Concepts

Index
(B⁺-tree)

Bistro Cafe ...

URLs

Vector Space Model for Relevance Ranking



Ranking by descending relevance

Search engine

Query $q \in [0,1]^{|F|}$ (set of weighted features)

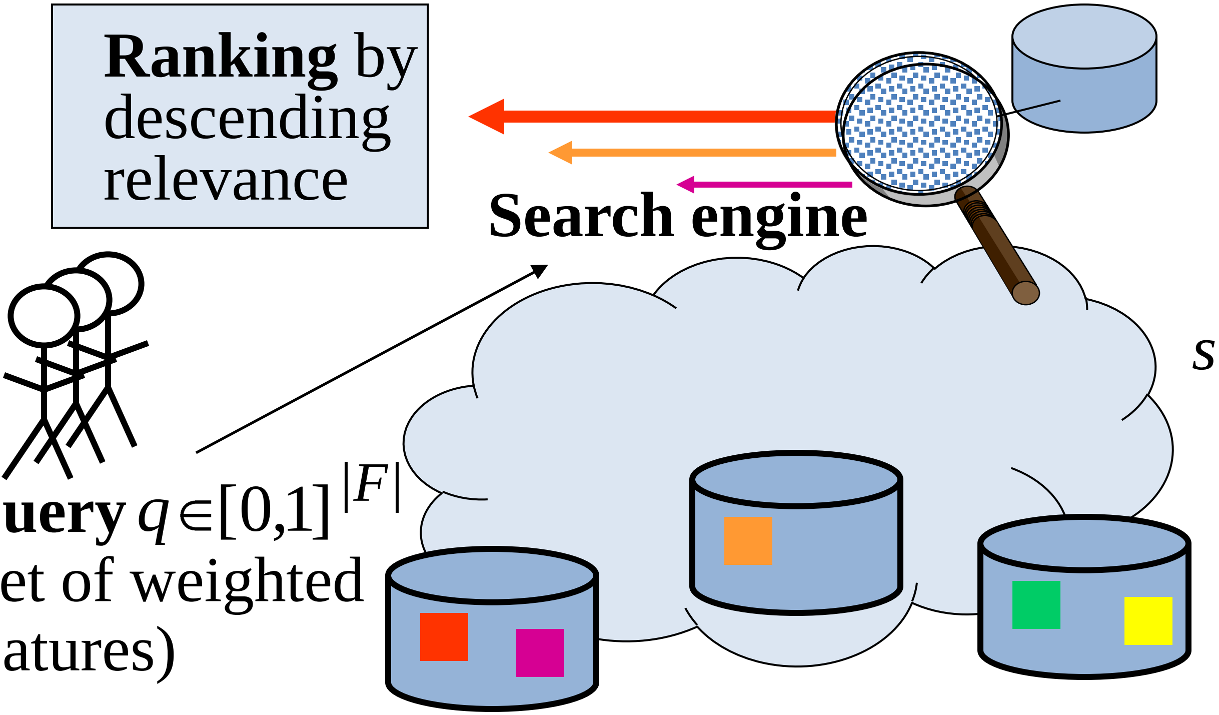
Documents are **feature vectors (bags of words)**

Similarity metric:

$$sim(d_i, q) := \frac{\sum_{j=1}^{|F|} d_{ij}\, q_j}{\sqrt{\sum_{j=1}^{|F|} d_{ij}^2 \sum_{j=1}^{|F|} q_j^2}}$$

$$with\ d_i \in [0,1]^{|F|}$$

Vector Space Model for Relevance Ranking



Ranking by descending relevance

Search engine

Query $q \in [0,1]^{|F|}$ (set of weighted features)

Documents are **feature vectors (bags of words)**

Similarity metric:

$$sim(d_i, q) := \frac{\sum_{j=1}^{|F|} d_{ij}\, q_j}{\sqrt{\sum_{j=1}^{|F|} d_{ij}^2 \sum_{j=1}^{|F|} q_j^2}}$$

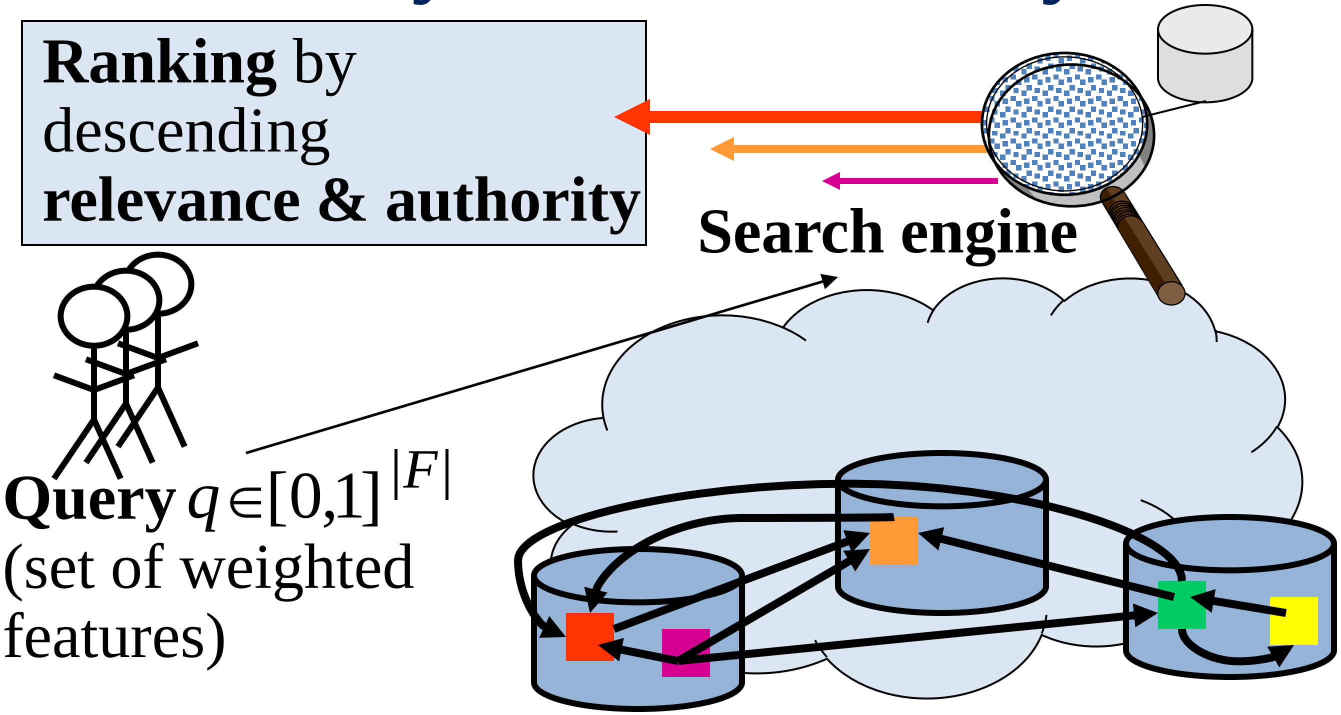
$$with\ d_i \in [0,1]^{|F|}$$

e.g., using: $d_{ij} := w_{ij} / \sqrt{\sum_k w_{ik}^2}$

$$w_{ij} := \log\left(1 + \frac{freq(f_j, d_i)}{\max_k\ freq(f_k, d_i)}\right) \log \frac{\#docs}{\#docs\ with\ f_i}$$

tf*idf formula

Link Analysis for Authority Ranking



Ranking by descending **relevance & authority**

Search engine

Query $q \in [0,1]^{|F|}$
(set of weighted features)

+ Consider in-degree and out-degree of Web nodes:
Authority Rank (d_i) :=
Stationary visitation probability [d_i]
in random walk on the Web (ergodic Markov Chain)

+ Reconciliation of relevance and authority by ad hoc weighting

Google's PageRank in a Nutshell [Page/Brin 1998]

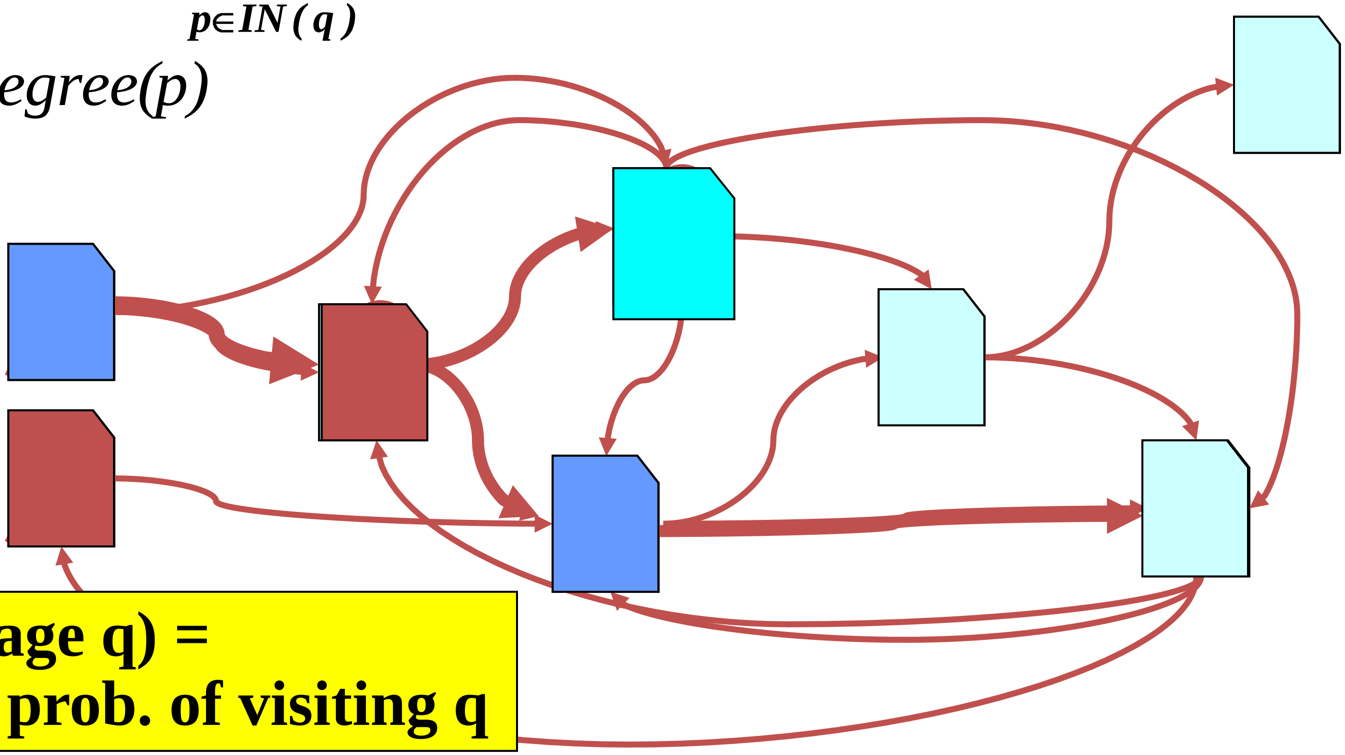
PageRank (PR): links are endorsements & increase page authority, authority is higher if links come from high-authority pages

$$PR(q) = \varepsilon \cdot j(q) + (1-\varepsilon) \cdot \sum_{p \in IN(q)} PR(p) \cdot t(p,q)$$

"Social" Ranking

with $t(p,q) = 1/outdegree(p)$

and $j(q) = 1/N$



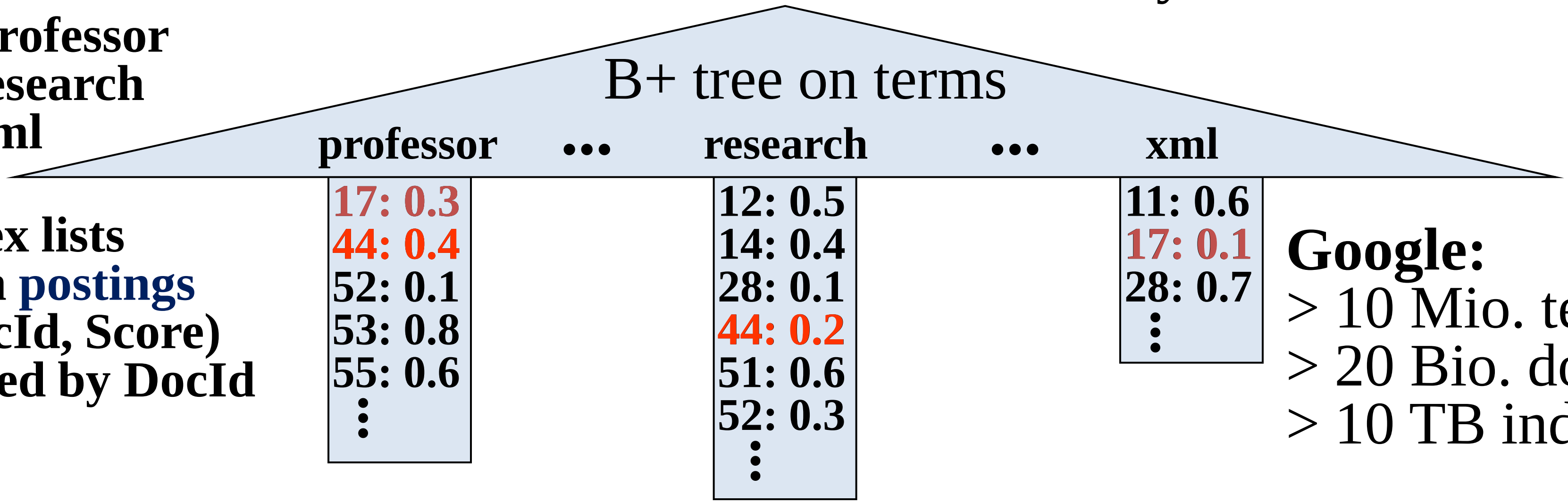
Authority (page q) = stationary prob. of visiting q

Random walk: uniform-randomly choose links & random jumps

Indexing with Inverted Lists

Vector space model suggests **term-document matrix**,
but data is sparse and queries are even very sparse
→ better use **inverted index lists** with terms as keys for B+ tree

q: professor
research
xml



B+ tree on terms

professor	•••	research	•••	xml
17: 0.3		12: 0.5		11: 0.6
44: 0.4		14: 0.4		17: 0.1
52: 0.1		28: 0.1		28: 0.7
53: 0.8		44: 0.2		⋮
55: 0.6		51: 0.6		
⋮		52: 0.3		
		⋮		

index lists
with postings
(DocId, Score)
sorted by DocId

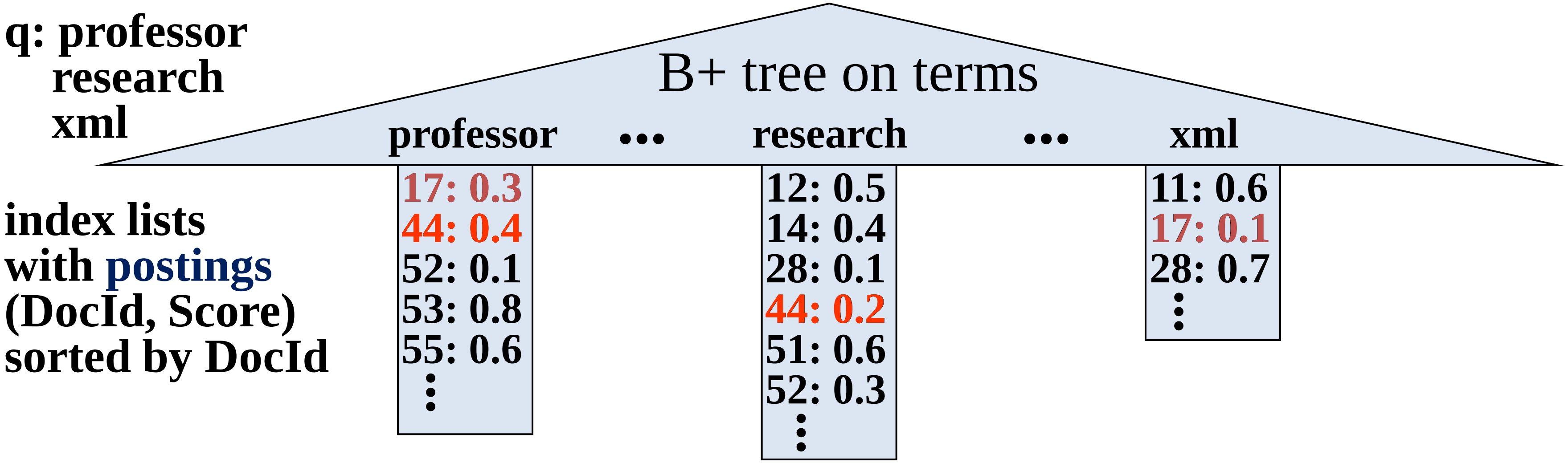
Google:
> 10 Mio. terms
> 20 Bio. docs
> 10 TB index

terms can be full words, word stems, word pairs, substrings, N-grams, etc.
(whatever "dictionary terms" we prefer for the application)

- index-list entries in **DocId order** for fast Boolean operations
- many techniques for excellent **compression** of index lists
- additional **position index** needed for phrases, proximity, etc.

(or other pre-computed data structures)

Query Processing on Inverted Lists



q: professor
research
xml

index lists with postings (DocId, Score) sorted by DocId

B+ tree on terms

professor	**research**	**xml**
17: 0.3	12: 0.5	11: 0.6
44: 0.4	14: 0.4	17: 0.1
52: 0.1	28: 0.1	28: 0.7
53: 0.8	44: 0.2	⋮
55: 0.6	51: 0.6	
⋮	52: 0.3	
	⋮	

Given: query $q = t_1\ t_2 \dots t_z$ with z (conjunctive) keywords
similarity scoring function ***score(q,d)*** for docs $d \in D$, e.g.: $\vec{q} \cdot \vec{d}$
with precomputed scores (index weights) $s_i(d)$ for which $q_i \neq 0$

Find: top-k results for ***score(q,d) =aggr{s_i(d)}*** (e.g.: $\Sigma_{i \in q}\ s_i(d)$)

Join-then-sort algorithm:

top-k (
$\sigma[term=t_1]\ (index) \bowtie_{DocId}$
$\sigma[term=t_2]\ (index) \bowtie_{DocId}$
$\dots \bowtie_{DocId}$
$\sigma[term=t_z]\ (index)$ order by s desc)

Evaluation of Search Result Quality: Basic Measures

Ideal measure is "**satisfaction of user's information need**"
heuristically approximated by benchmarking measures
(on test corpora with query suite and relevance assessment by experts)

Capability to return **only** relevant documents:

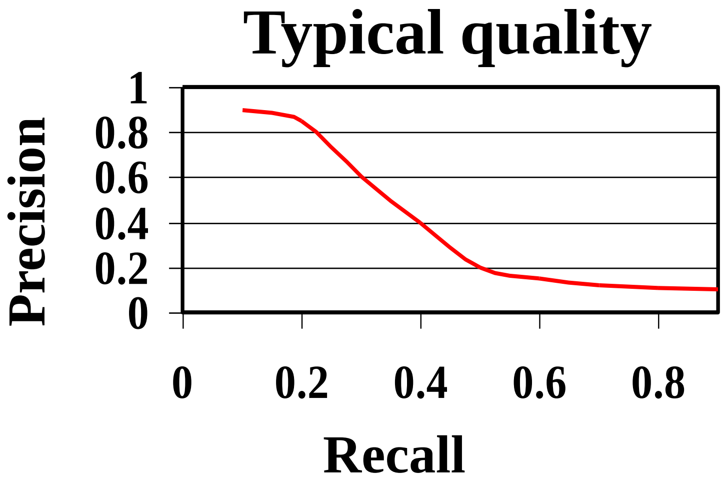
$$\textbf{\textit{Precision}} = \frac{\textit{\# relevant docs among top r}}{r}$$

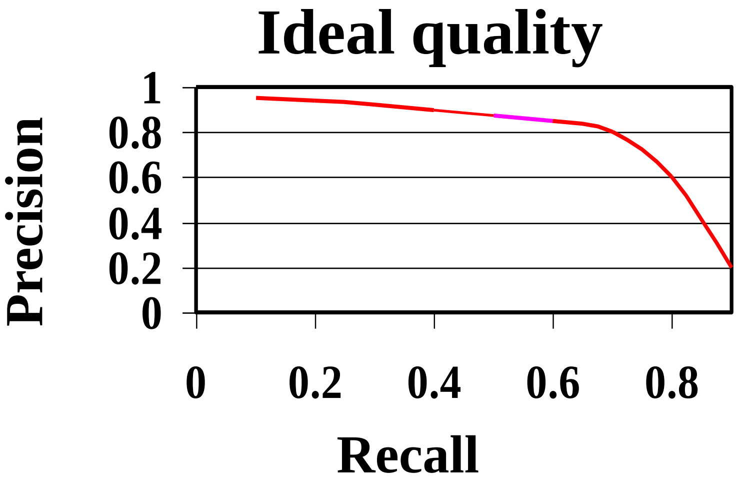
typically for
r = 10, 100, 1000

Capability to return **all** relevant documents:

$$\textbf{\textit{Recall}} = \frac{\textit{\# relevant docs among top r}}{\textit{\# relevant docs}}$$

typically for
r = corpus size





Deep Web (Hidden Web)

Data (in DBS or CMS) accessible only through query interfaces:
HTML forms, API's (e.g. Web Services with WSDL/REST)

Study by B. He, M. Patel, Z. Zhang, K. Chang, CACM 2006:
> 300 000 sites with > 450 000 databases and > 1 200 000 interfaces
coverage in directories (e.g. dmoz.org) is < 15%,
total data volume estimated **10-100 PBytes**

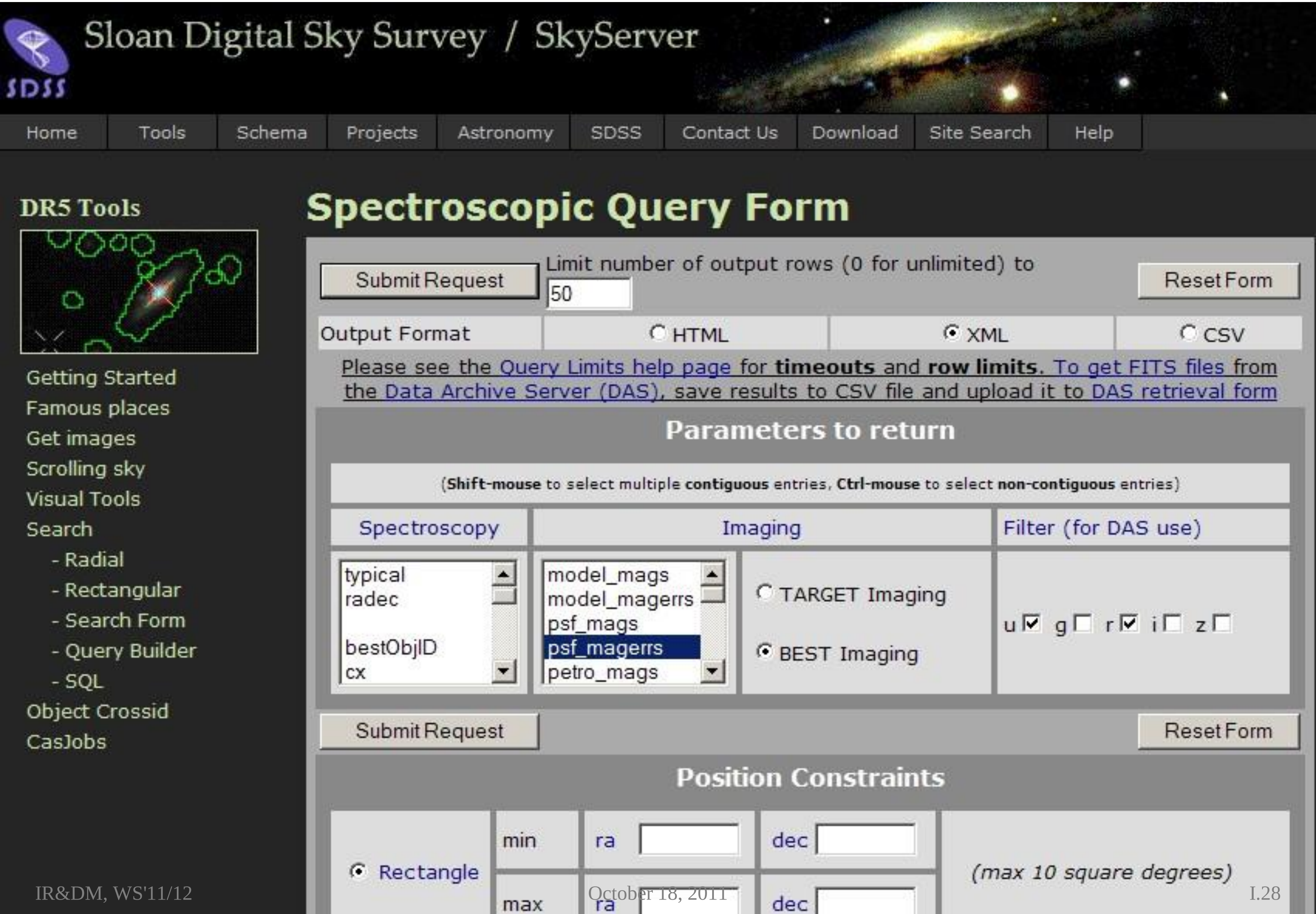
Examples of Deep Web sources:

e-business and entertainment: amazon.com, ebay.com, realtor.com, cars.com, imdb.com, reviews-zdnet.com, epinions.com

news, libraries, society: cnn.com, yahoo.com, spiegel.de, deutschland.de, uspto.gov, loc.gov, dip.bundestag.de, destatis.de, ddb.de, bnf.fr, kb.nl, kb.se, weatherimages.org, TerraServer.com, lonelyplanet.com

e-science: NCBI, SRS, SwissProt, PubMed, SkyServer, GriPhyN

Example SkyServer http://skyserver.sdss.org



Spectroscopic Query Form

Limit number of output rows (0 for unlimited) to 50

Output Format ○ HTML ● XML ○ CSV

Please see the Query Limits help page for **timeouts** and **row limits**. To get FITS files from the Data Archive Server (DAS), save results to CSV file and upload it to DAS retrieval form

Parameters to return

(**Shift-mouse** to select multiple **contiguous** entries, **Ctrl-mouse** to select **non-contiguous** entries)

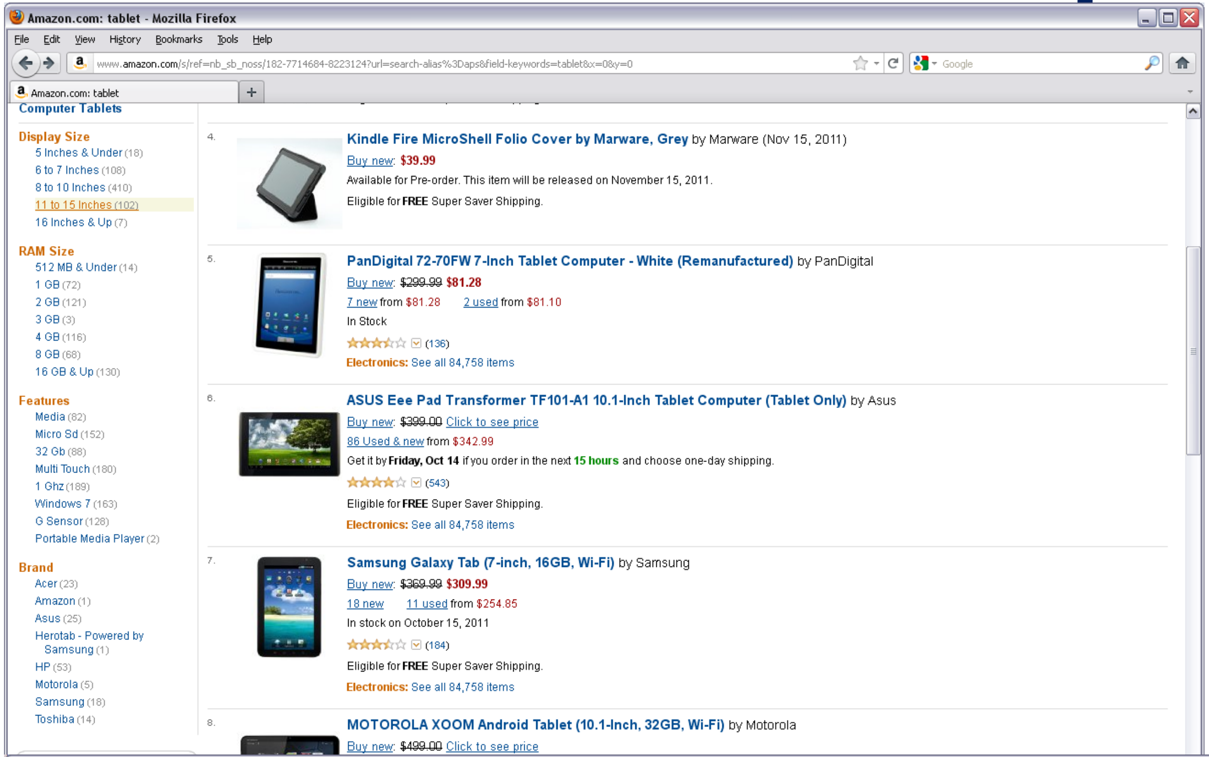
Spectroscopy	Imaging		Filter (for DAS use)
typical radec bestObjID cx	model_mags model_magerrs psf_mags psf_magerrs petro_mags	○ TARGET Imaging ● BEST Imaging	u ☑ g □ r ☑ i □ z □

Position Constraints

● Rectangle	min	ra	dec
	max	ra	dec

(max 10 square degrees)

Faceted Search on Deep-Web Sources



Amazon.com: tablet - Mozilla Firefox

File Edit View History Bookmarks Tools Help

www.amazon.com/s/ref=nb_sb_noss/182-7714684-8223124?url=search-alias%3Daps&field-keywords=tablet&x=0&y=0

Amazon.com: tablet

Computer Tablets

Display Size
- 5 Inches & Under (18)
- 6 to 7 Inches (108)
- 8 to 10 Inches (410)
- 11 to 15 Inches (102)
- 16 Inches & Up (7)

RAM Size
- 512 MB & Under (14)
- 1 GB (72)
- 2 GB (121)
- 3 GB (3)
- 4 GB (116)
- 8 GB (68)
- 16 GB & Up (130)

Features
- Media (82)
- Micro Sd (152)
- 32 Gb (88)
- Multi Touch (180)
- 1 Ghz (189)
- Windows 7 (163)
- G Sensor (128)
- Portable Media Player (2)

Brand
- Acer (23)
- Amazon (1)
- Asus (25)
- Herotab - Powered by Samsung (1)
- HP (53)
- Motorola (5)
- Samsung (18)
- Toshiba (14)

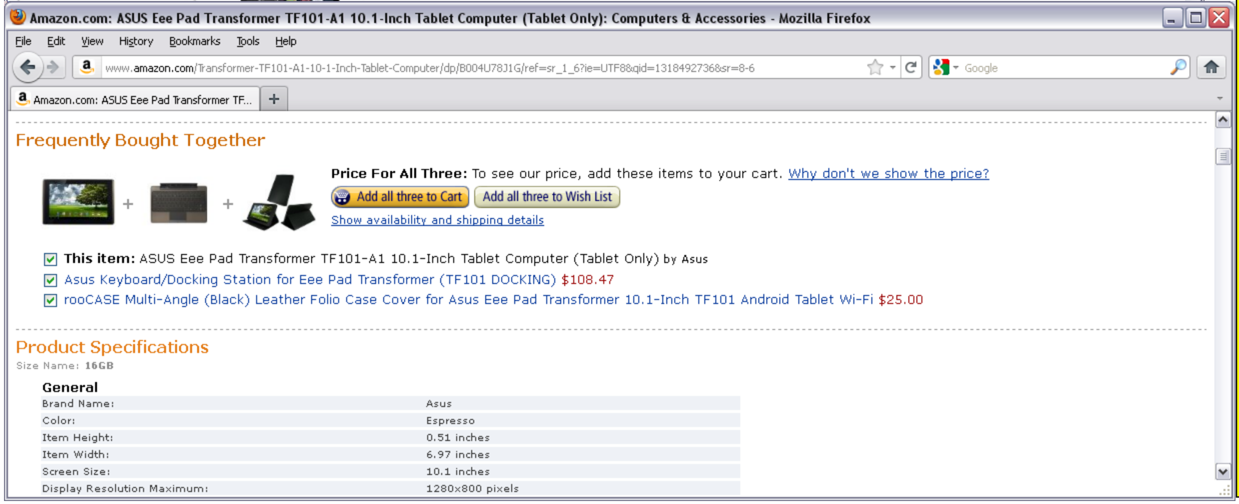
4. Kindle Fire MicroShell Folio Cover by Marware, Grey by Marware (Nov 15, 2011)
Buy new: $39.99
Available for Pre-order. This item will be released on November 15, 2011.
Eligible for **FREE** Super Saver Shipping.

5. PanDigital 72-70FW 7-Inch Tablet Computer - White (Remanufactured) by PanDigital
Buy new: ~~$299.99~~ $81.28
7 new from $81.28 2 used from $81.10
In Stock
(136)
Electronics: See all 84,758 items

6. ASUS Eee Pad Transformer TF101-A1 10.1-Inch Tablet Computer (Tablet Only) by Asus
Buy new: ~~$399.00~~ Click to see price
86 Used & new from $342.99
Get it by **Friday, Oct 14** if you order in the next **15 hours** and choose one-day shipping.
(543)
Eligible for **FREE** Super Saver Shipping.
Electronics: See all 84,758 items

7. Samsung Galaxy Tab (7-inch, 16GB, Wi-Fi) by Samsung
Buy new: ~~$369.99~~ $309.99
18 new 11 used from $254.85
In stock on October 15, 2011
(184)
Eligible for **FREE** Super Saver Shipping.
Electronics: See all 84,758 items

8. MOTOROLA XOOM Android Tablet (10.1-Inch, 32GB, Wi-Fi) by Motorola
Buy new: ~~$499.00~~ Click to see price



Amazon.com: ASUS Eee Pad Transformer TF101-A1 10.1-Inch Tablet Computer (Tablet Only): Computers & Accessories - Mozilla Firefox

File Edit View History Bookmarks Tools Help

www.amazon.com/Transformer-TF101-A1-10-1-Inch-Tablet-Computer/dp/B004U78J1G/ref=sr_1_6?ie=UTF8&qid=1318492736&sr=8-6

Amazon.com: ASUS Eee Pad Transformer TF...

Frequently Bought Together

Price For All Three: To see our price, add these items to your cart. Why don't we show the price?

Add all three to Cart | Add all three to Wish List

Show availability and shipping details

☑ **This item:** ASUS Eee Pad Transformer TF101-A1 10.1-Inch Tablet Computer (Tablet Only) by Asus
☑ Asus Keyboard/Docking Station for Eee Pad Transformer (TF101 DOCKING) $108.47
☑ rooCASE Multi-Angle (Black) Leather Folio Case Cover for Asus Eee Pad Transformer 10.1-Inch TF101 Android Tablet Wi-Fi $25.00

Product Specifications

Size Name: **16GB**

General

Brand Name:	Asus
Color:	Espresso
Item Height:	0.51 inches
Item Width:	6.97 inches
Screen Size:	10.1 inches
Display Resolution Maximum:	1280x800 pixels

- Products grouped by **facets** (characteristic properties)
- Facets form **lattices**
  - Drill-down
  - Roll-up
- **Classical data-mining example:**

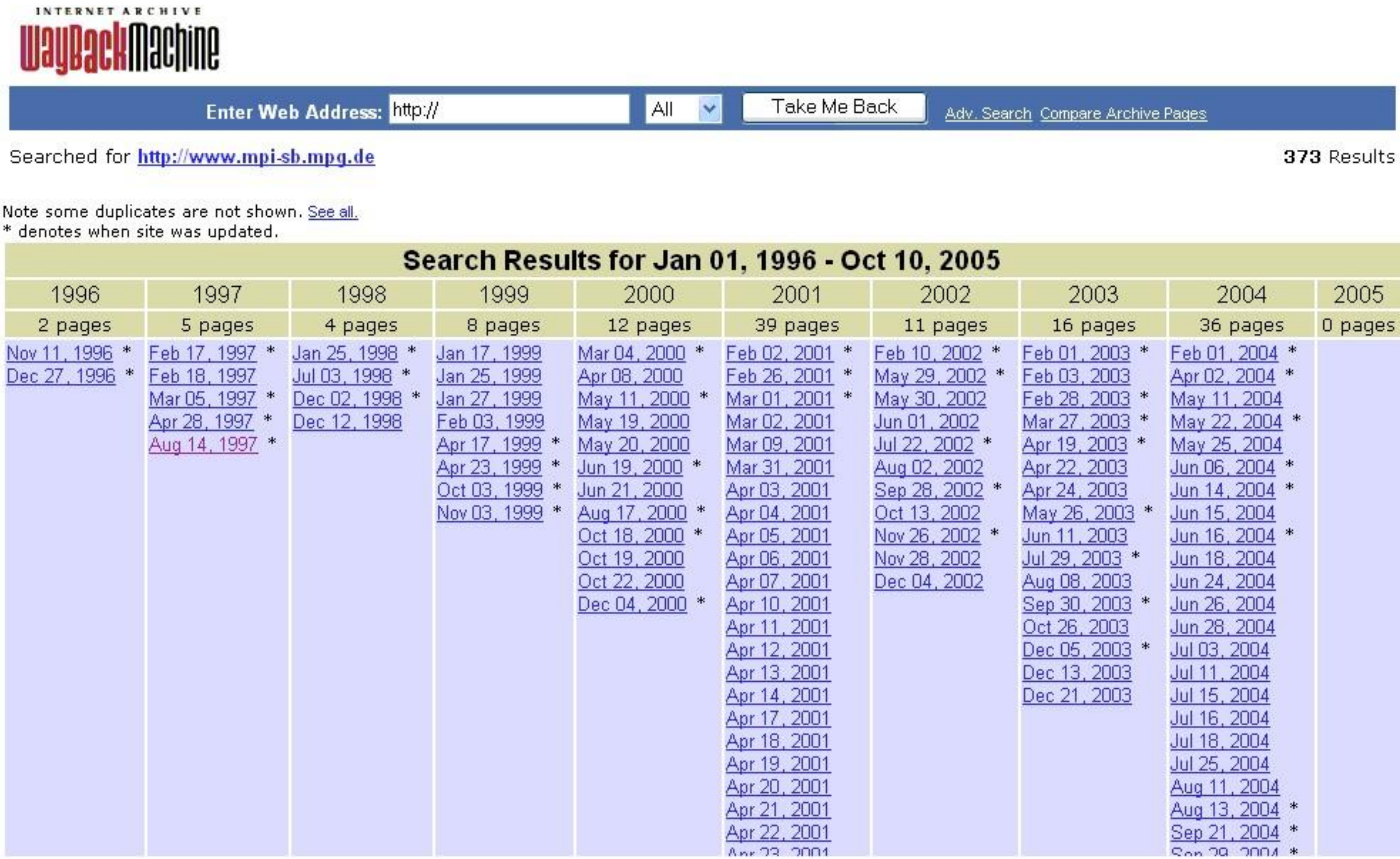
  “Other user who bought this item also bought …”

→ **Frequent item sets**

→ **“Basket Mining”**

Web Archiving

http://www.archive.org



Enter Web Address: http:// All

Adv. Search Compare Archive Pages

Searched for http://www.mpi-sb.mpg.de

373 Results

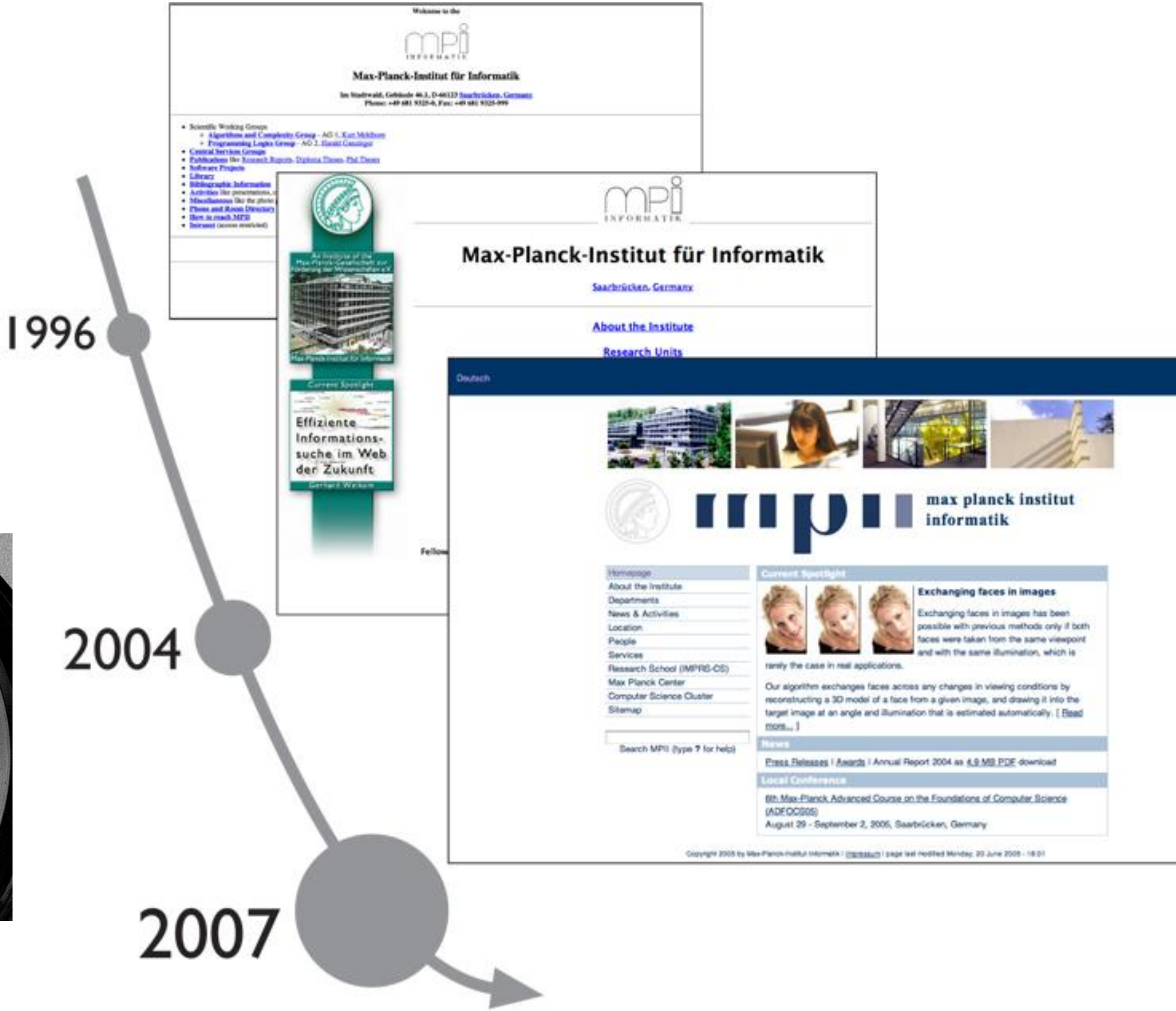
Note some duplicates are not shown. See all.
* denotes when site was updated.

Search Results for Jan 01, 1996 - Oct 10, 2005

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
2 pages	5 pages	4 pages	8 pages	12 pages	39 pages	11 pages	16 pages	36 pages	0 pages
Nov 11, 1996 *	Feb 17, 1997 *	Jan 25, 1998 *	Jan 17, 1999	Mar 04, 2000 *	Feb 02, 2001 *	Feb 10, 2002 *	Feb 01, 2003 *	Feb 01, 2004 *	
Dec 27, 1996 *	Feb 18, 1997	Jul 03, 1998 *	Jan 25, 1999	Apr 08, 2000	Feb 26, 2001 *	May 29, 2002 *	Feb 03, 2003	Apr 02, 2004 *	
	Mar 05, 1997 *	Dec 02, 1998 *	Jan 27, 1999	May 11, 2000 *	Mar 01, 2001 *	May 30, 2002	Feb 28, 2003 *	May 11, 2004	
	Apr 28, 1997 *	Dec 12, 1998	Feb 03, 1999	May 19, 2000	Mar 02, 2001	Jun 01, 2002	Mar 27, 2003 *	May 22, 2004 *	
	Aug 14, 1997 *		Apr 17, 1999 *	May 20, 2000	Mar 09, 2001	Jul 22, 2002 *	Apr 19, 2003 *	May 25, 2004	
			Apr 23, 1999 *	Jun 19, 2000 *	Mar 31, 2001	Aug 02, 2002	Apr 22, 2003	Jun 06, 2004 *	
			Oct 03, 1999 *	Jun 21, 2000	Apr 03, 2001	Sep 28, 2002 *	Apr 24, 2003	Jun 14, 2004 *	
			Nov 03, 1999 *	Aug 17, 2000 *	Apr 04, 2001	Oct 13, 2002	May 26, 2003 *	Jun 15, 2004	
				Oct 18, 2000 *	Apr 05, 2001	Nov 26, 2002 *	Jun 11, 2003	Jun 16, 2004 *	
				Oct 19, 2000	Apr 06, 2001	Nov 28, 2002	Jul 29, 2003 *	Jun 18, 2004	
				Oct 22, 2000	Apr 07, 2001	Dec 04, 2002	Aug 08, 2003	Jun 24, 2004	
				Dec 04, 2000 *	Apr 10, 2001		Sep 30, 2003 *	Jun 26, 2004	
					Apr 11, 2001		Oct 26, 2003	Jun 28, 2004	
					Apr 12, 2001		Dec 05, 2003 *	Jul 03, 2004	
					Apr 13, 2001		Dec 13, 2003	Jul 11, 2004	
					Apr 14, 2001		Dec 21, 2003	Jul 15, 2004	
					Apr 17, 2001			Jul 16, 2004	
					Apr 18, 2001			Jul 18, 2004	
					Apr 19, 2001			Jul 25, 2004	
					Apr 20, 2001			Aug 11, 2004	
					Apr 21, 2001			Aug 13, 2004 *	
					Apr 22, 2001			Sep 21, 2004 *	
					Apr 23, 2001			Sep 29, 2004 *	

40 Billion URLs archived every 2 months since 1996 → **500 TBytes**

Time Travel in Web Archives



Welcome to the

Max-Planck-Institut für Informatik

Im Stadtwald, Gebäude 46.1, D-66123 Saarbrücken, Germany
Phone: +49 681 9325-0, Fax: +49 681 9325-999

- Scientific Working Groups
  - Algorithms and Complexity Group - AG 1, Kurt Mehlhorn
  - Programming Logics Group - AG 2, Harald Ganzinger
- Central Services Groups
- Publications like Research Reports, Diploma Theses, Phd Theses
- Software Projects
- Library
- Bibliographic Information
- Activities like presentations,
- Miscellaneous like the photo
- Phone and Room Directory
- How to reach MPII
- Intranet (access restricted)

An Institute of the Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V.

Max-Planck-Institut für Informatik

Current Spotlight

Effiziente Informationssuche im Web der Zukunft

Gerhard Weikum

Max-Planck-Institut für Informatik

Saarbrücken, Germany

About the Institute

Research Units

Fellow

Deutsch

max planck institut informatik

- Homepage
- About the Institute
- Departments
- News & Activities
- Location
- People
- Services
- Research School (IMPRS-CS)
- Max Planck Center
- Computer Science Cluster
- Sitemap

Search MPII (type ? for help)

Current Spotlight

Exchanging faces in images

Exchanging faces in images has been possible with previous methods only if both faces were taken from the same viewpoint and with the same illumination, which is rarely the case in real applications.

Our algorithm exchanges faces across any changes in viewing conditions by reconstructing a 3D model of a face from a given image, and drawing it into the target image at an angle and illumination that is estimated automatically. [Read more...]

News

Press Releases | Awards | Annual Report 2004 as 4.9 MB PDF download

Local Conference

6th Max-Planck Advanced Course on the Foundations of Computer Science (ADFOCS05)
August 29 - September 2, 2005, Saarbrücken, Germany

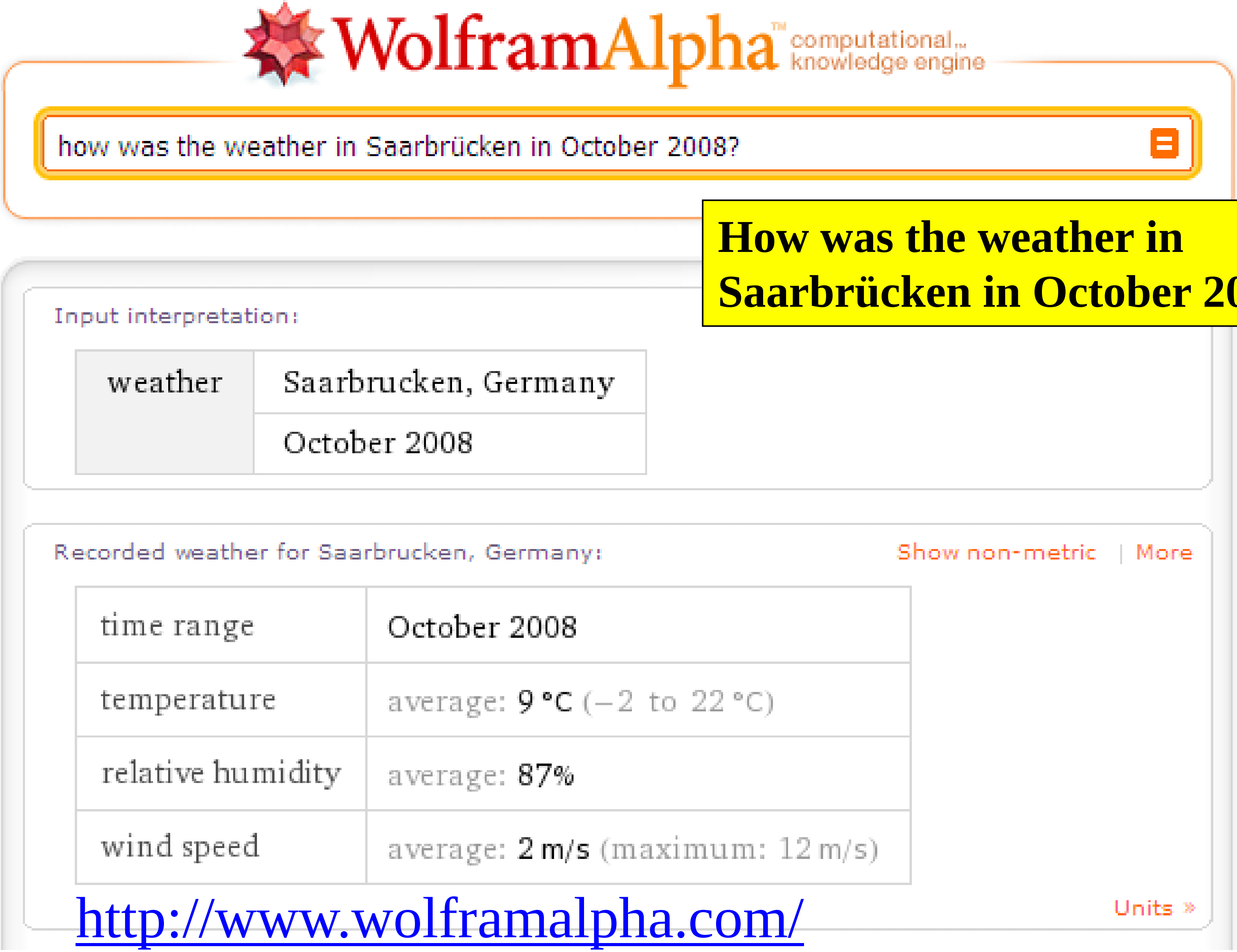
Copyright 2005 by Max-Planck-Institut Informatik | Impressum | page last modified Monday, 20 June 2005 - 18:01

Beyond Google: Search for Knowledge

Answer “knowledge queries” **(by scientists, journalists, analysts, etc.)** **such as:**

- drugs or enzymes that inhibit proteases (HIV)
- German Nobel prize winner who survived both world wars and outlived all of his four children
- who was German chancellor when Angela Merkel was born
- how are Max Planck, Angela Merkel, and the Dalai Lama related
- politicians who are also scientists

Example: WolframAlpha



how was the weather in Saarbrücken in October 2008?

Input interpretation:

weather	Saarbrucken, Germany
	October 2008

Recorded weather for Saarbrucken, Germany:

time range	October 2008
temperature	average: 9 °C (–2 to 22 °C)
relative humidity	average: 87%
wind speed	average: 2 m/s (maximum: 12 m/s)

How was the weather in Saarbrücken in October 2008?

http://www.wolframalpha.com/

Semantic Search

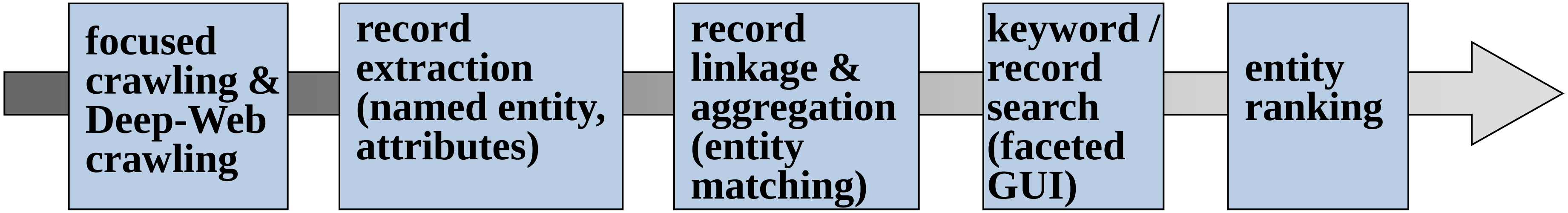
Search on **entities**, **attributes**, and **relationships**

→ focus on **structured data** sources (relational, XML, RDF)
→ leverage manually **annotated data** (social tagging, Web2.0)
→ perform **info extraction** on semi-structured & textual data

Motivation and Applications:

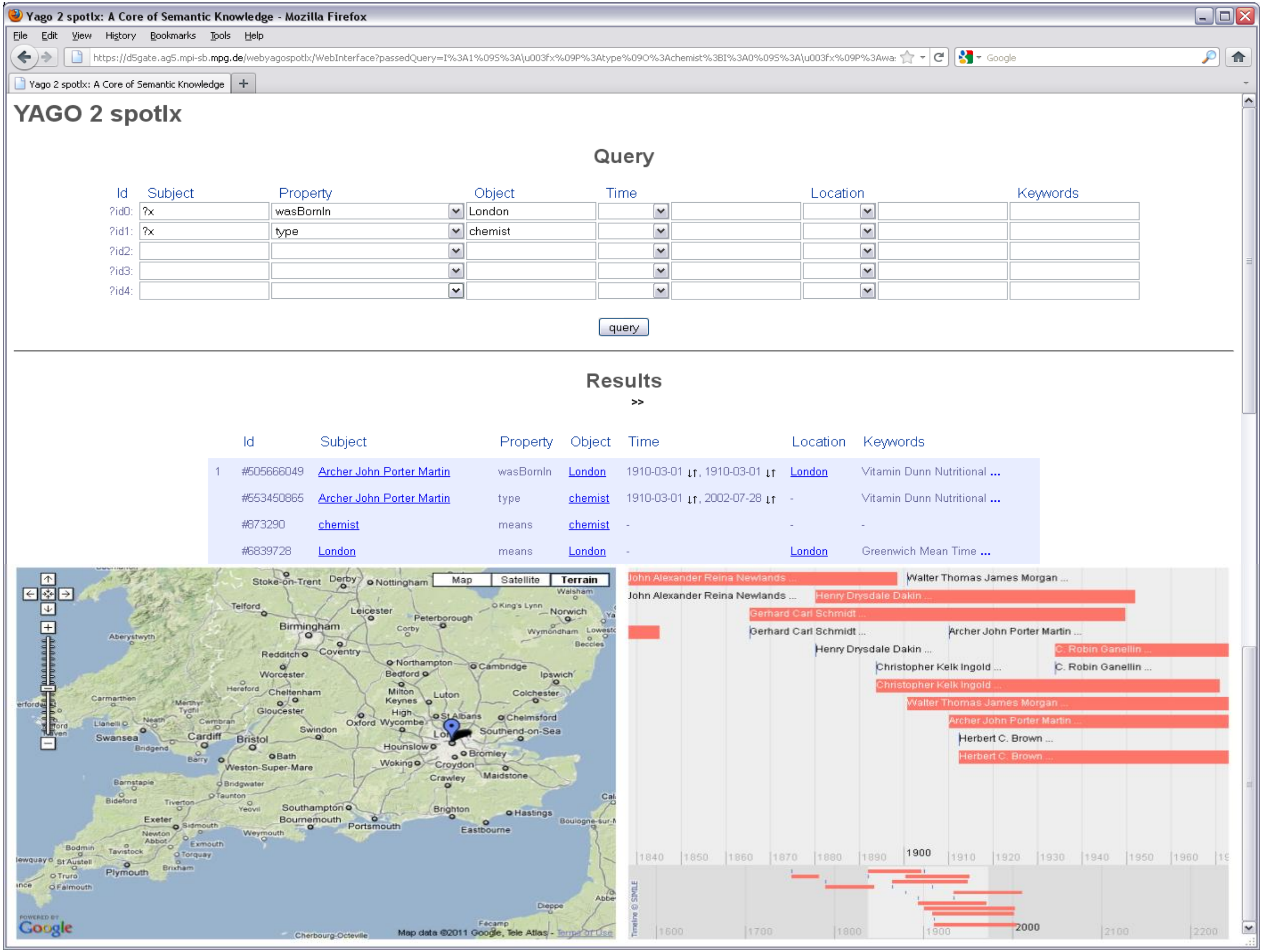
- Web search for vertical domains
  (products, traveling, entertainment, scholarly publications, etc.)
- backend for natural-language QA
- towards better Deep-Web search, digital libraries, e-science

System architecture:


focused crawling & Deep-Web crawling
record extraction (named entity, attributes)
record linkage & aggregation (entity matching)
keyword / record search (faceted GUI)
entity ranking

Example: YAGO-NAGA

http://www.mpi-inf.mpg.de/yago-naga/



Yago 2 spotlx: A Core of Semantic Knowledge - Mozilla Firefox

https://d5gate.ag5.mpi-sb.mpg.de/webyagospotlx/WebInterface?passedQuery=I%3A1%09S%3A\u003fx%09P%3Atype%09O%3Achemist%3BI%3A0%09S%3A\u003fx%09P%3Awa:

YAGO 2 spotlx

Query

Id	Subject	Property	Object	Time		Location		Keywords
?id0:	?x	wasBornIn	London					
?id1:	?x	type	chemist					
?id2:								
?id3:								
?id4:								

query

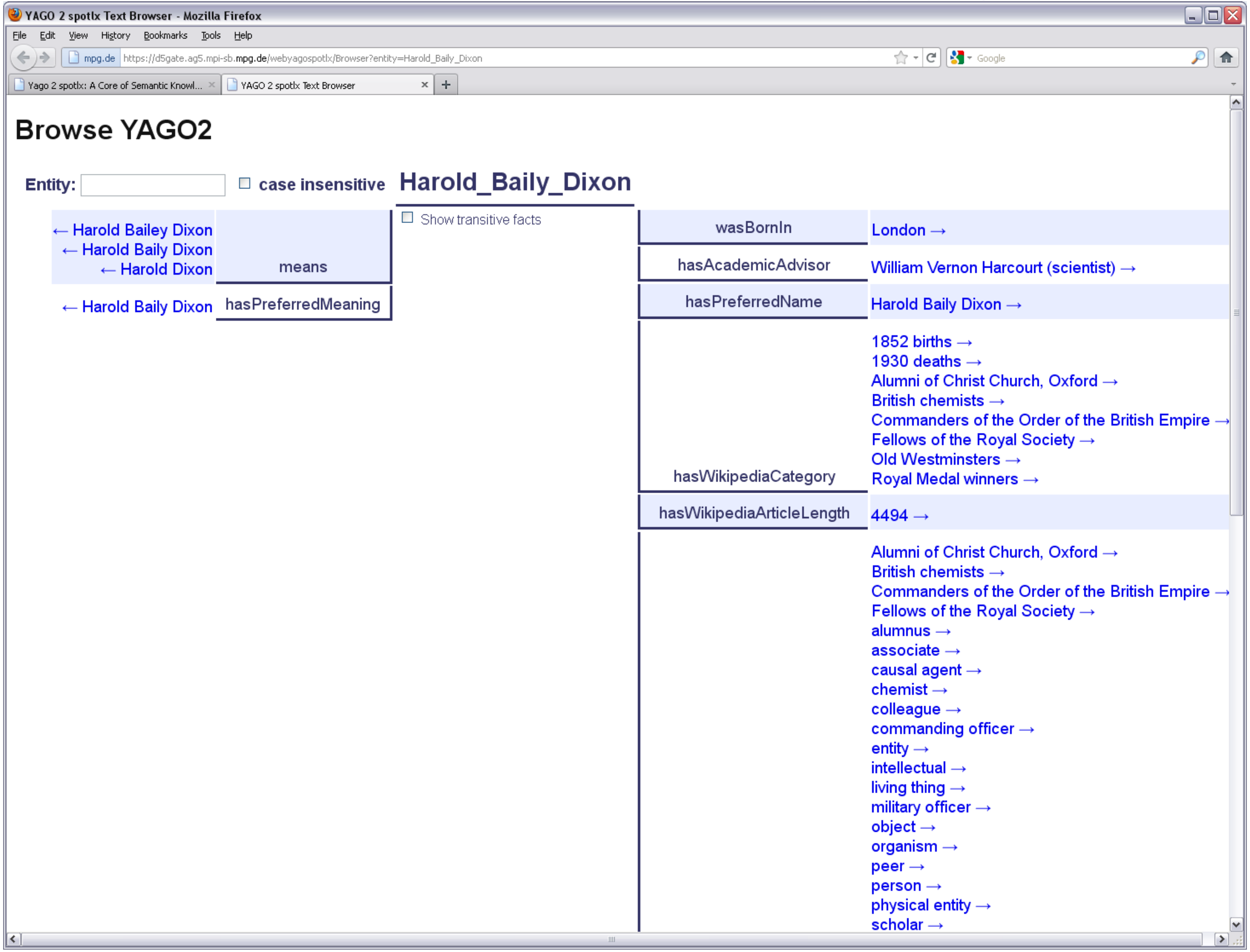
Results

>>

	Id	Subject	Property	Object	Time	Location	Keywords
1	#505666049	Archer John Porter Martin	wasBornIn	London	1910-03-01 ↓↑, 1910-03-01 ↓↑	London	Vitamin Dunn Nutritional ...
	#553450865	Archer John Porter Martin	type	chemist	1910-03-01 ↓↑, 2002-07-28 ↓↑	-	Vitamin Dunn Nutritional ...
	#873290	chemist	means	chemist	-	-	-
	#6839728	London	means	London	-	London	Greenwich Mean Time ...

Example: YAGO-NAGA

http://www.mpi-inf.mpg.de/yago-naga/


YAGO 2 spotlx Text Browser - Mozilla Firefox
File Edit View History Bookmarks Tools Help
mpg.de https://d5gate.ag5.mpi-sb.mpg.de/webyagospotlx/Browser?entity=Harold_Baily_Dixon
Google
Yago 2 spotlx: A Core of Semantic Knowl...
YAGO 2 spotlx Text Browser

Browse YAGO2

Entity: ☐ case insensitive **Harold_Baily_Dixon**

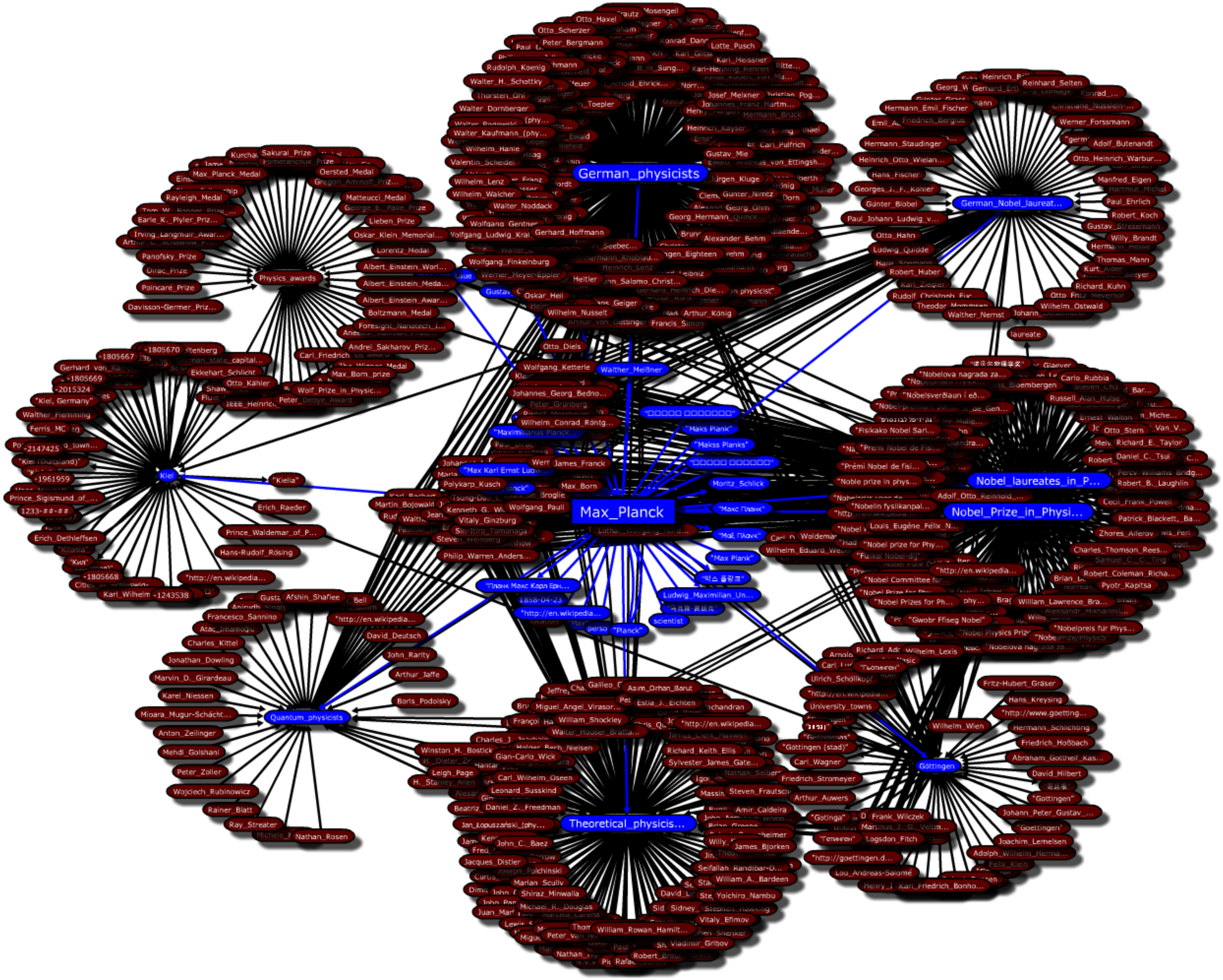
☐ Show transitive facts

← Harold Bailey Dixon ← Harold Baily Dixon ← Harold Dixon	means
← Harold Baily Dixon	hasPreferredMeaning

wasBornIn	London →
hasAcademicAdvisor	William Vernon Harcourt (scientist) →
hasPreferredName	Harold Baily Dixon →
hasWikipediaCategory	1852 births → 1930 deaths → Alumni of Christ Church, Oxford → British chemists → Commanders of the Order of the British Empire → Fellows of the Royal Society → Old Westminsters → Royal Medal winners →
hasWikipediaArticleLength	4494 →
	Alumni of Christ Church, Oxford → British chemists → Commanders of the Order of the British Empire → Fellows of the Royal Society → alumnus → associate → causal agent → chemist → colleague → commanding officer → entity → intellectual → living thing → military officer → object → organism → peer → person → physical entity → scholar →

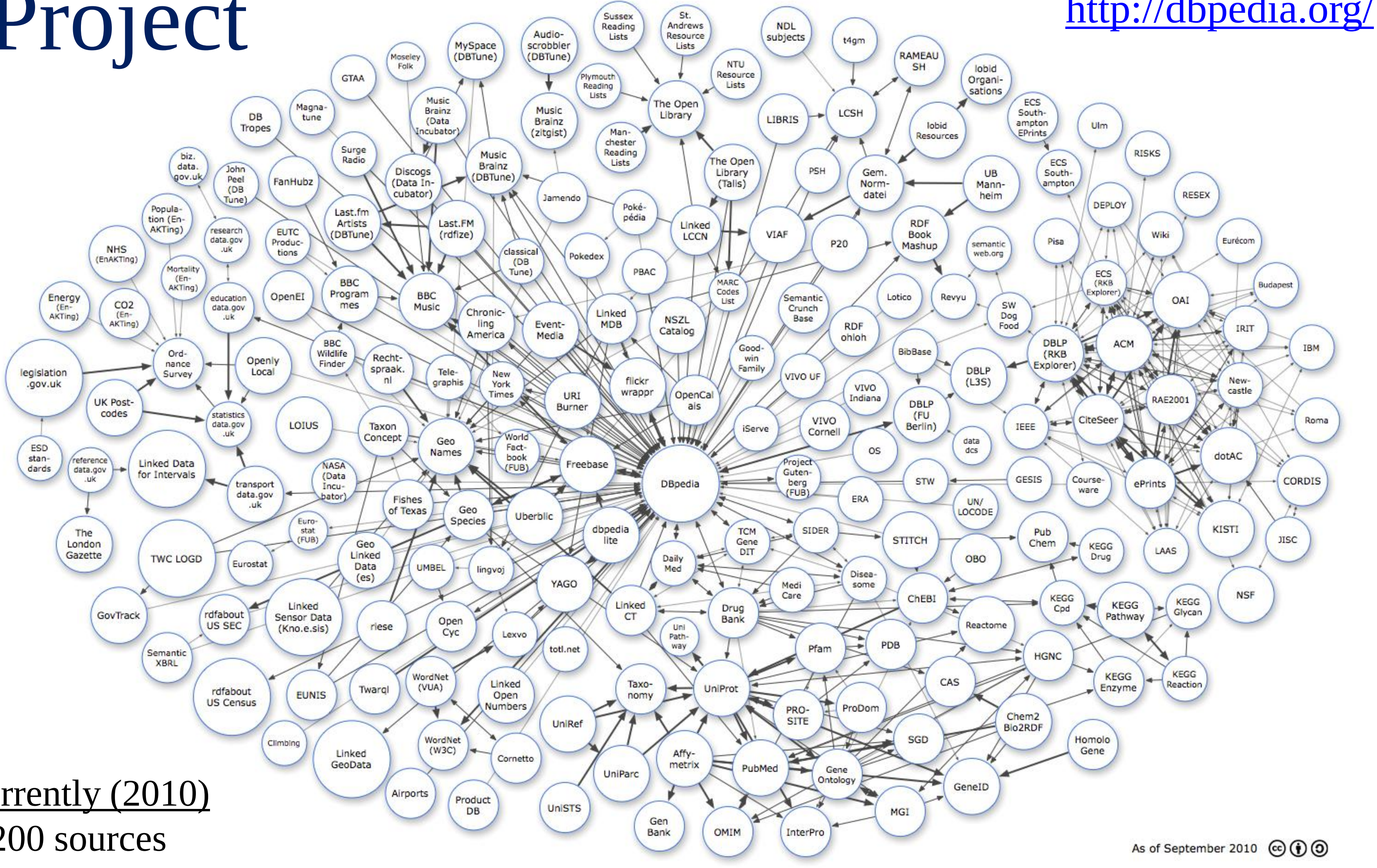
Example: URDF

http://urdf.mpi-inf.mpg.de/



The Linking Open Data (LOD) Project

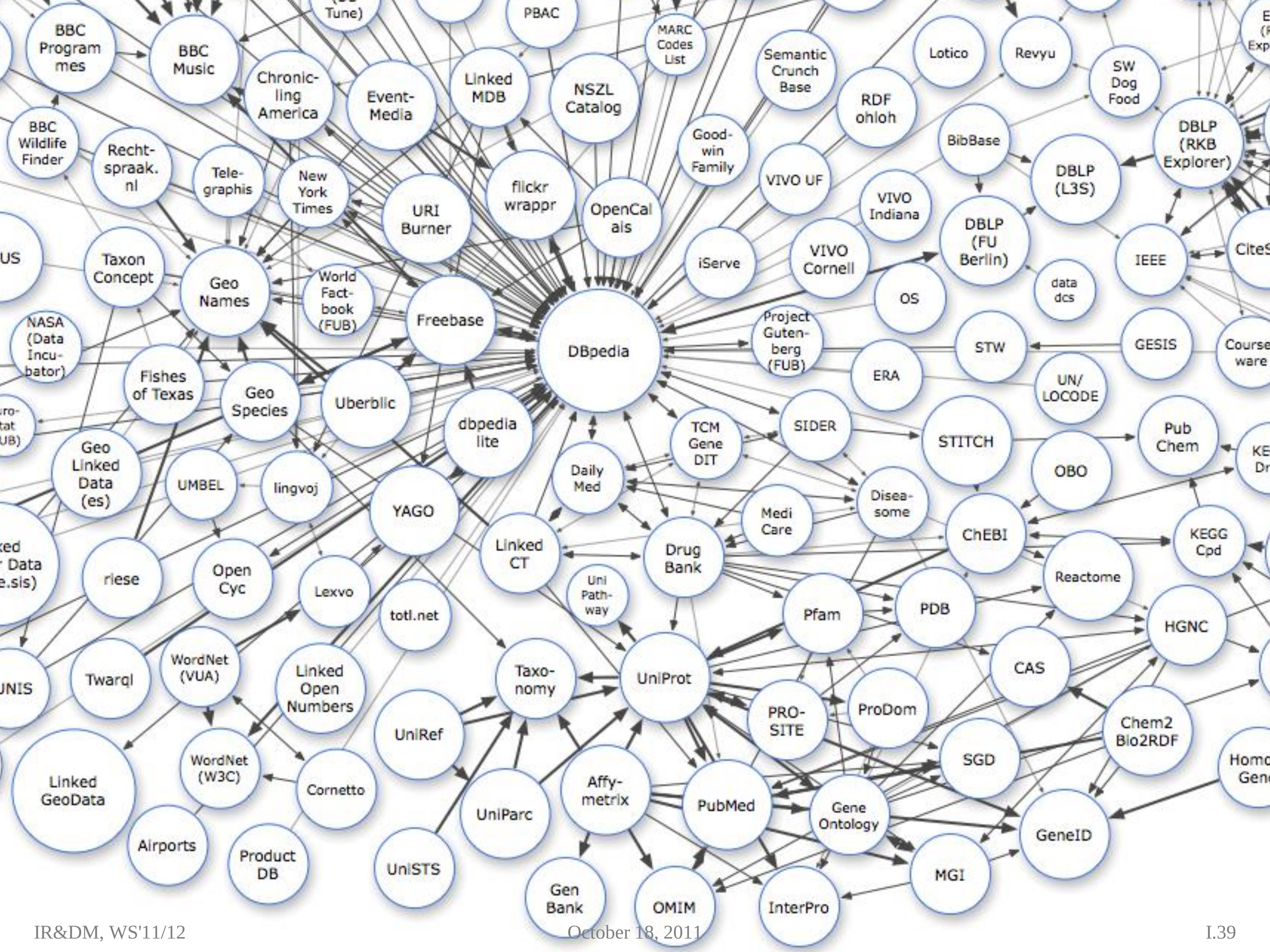
http://dbpedia.org/



<u>Currently (2010)</u>

- 200 sources
- 25 billion triples
- 400 million links

http://richard.cyganiak.de/2007/10/lod/imagemap.html


BBC Programmes
BBC Music
Chronicling America
Event-Media
Linked MDB
PBAC
MARC Codes List
NSZL Catalog
Semantic Crunch Base
Lotico
Revyu
SW Dog Food
RDF ohloh
DBLP (RKB Explorer)
BBC Wildlife Finder
Recht-spraak.nl
Tele-graphis
New York Times
URI Burner
flickr wrappr
OpenCalais
Good-win Family
VIVO UF
VIVO Indiana
BibBase
DBLP (L3S)
DBLP (FU Berlin)
IEEE
US
Taxon Concept
Geo Names
World Fact-book (FUB)
Freebase
DBpedia
iServe
VIVO Cornell
OS
data dcs
NASA (Data Incu-bator)
Project Guten-berg (FUB)
STW
GESIS
Course ware
ERA
UN/LOCODE
Fishes of Texas
Geo Species
Uberblic
dbpedia lite
TCM Gene DIT
SIDER
STITCH
Pub Chem
Geo Linked Data (es)
UMBEL
lingvoj
YAGO
Daily Med
Medi Care
Disea-some
OBO
ChEBI
KEGG Cpd
riese
Open Cyc
Lexvo
Linked CT
Drug Bank
Reactome
Uni Path-way
totl.net
Pfam
PDB
HGNC
Twarql
WordNet (VUA)
Linked Open Numbers
Taxo-nomy
UniProt
CAS
PRO-SITE
ProDom
Chem2 Bio2RDF
UniRef
SGD
Linked GeoData
WordNet (W3C)
Cornetto
Affy-metrix
UniParc
PubMed
Gene Ontology
GeneID
Airports
Product DB
UniSTS
Gen Bank
OMIM
InterPro
MGI

Multimodal Web (Images, Videos, NLP, …)

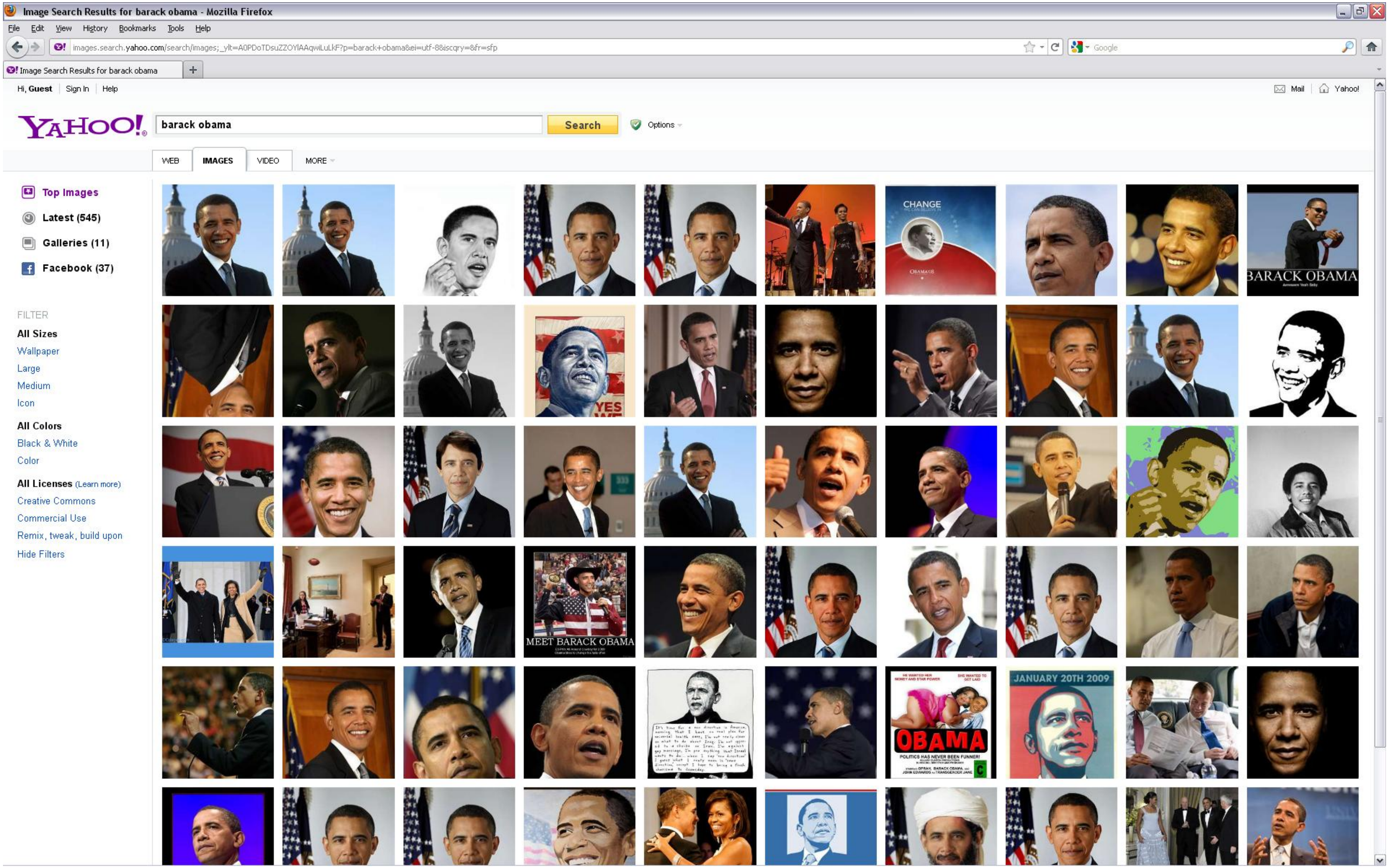
Search for **images**, **speech**, **audio files**, **videos**, etc.:

- based on **signal-level content features** (color distribution, contours, textures, video shot sequence, pitch change patterns, harmonic and rythmic features, etc. etc.)
- complement signal-level features with **annotations** from context (e.g. adjacent text in Web page, GPS coordinates from digital camera)
- **query by example**: similarity search w.r.t. given object(s) plus relevance feedback

Question answering (QA) in natural language:

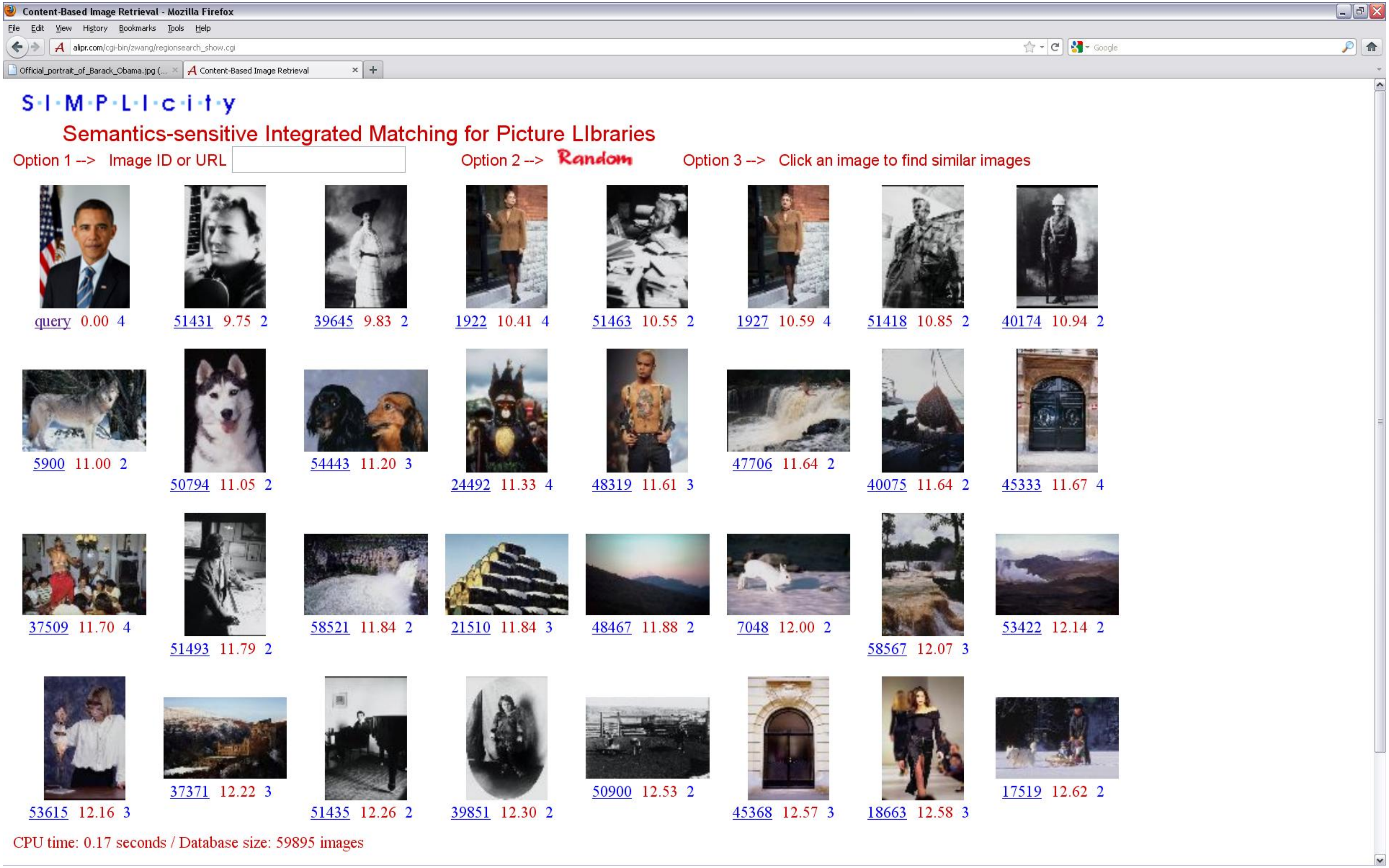
- express query as NL question: Who ..., When ..., Where ..., What ...
- provide short NL passages as query result(s), not entire documents

Internet Image Search



http://www.bing.com/images/

Content-based Image Retrieval by Example


Content-Based Image Retrieval - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
Google
Official_portrait_of_Barack_Obama.jpg (...
Content-Based Image Retrieval

alipr.com/cgi-bin/zwang/regionsearch_show.cgi

S·I·M·P·L·I·c·i·t·y

Semantics-sensitive Integrated Matching for Picture LIbraries

Option 1 --> Image ID or URL

Option 2 --> Random

Option 3 --> Click an image to find similar images

query 0.00 4
51431 9.75 2
39645 9.83 2
1922 10.41 4
51463 10.55 2
1927 10.59 4
51418 10.85 2
40174 10.94 2

5900 11.00 2
50794 11.05 2
54443 11.20 3
24492 11.33 4
48319 11.61 3
47706 11.64 2
40075 11.64 2
45333 11.67 4

37509 11.70 4
51493 11.79 2
58521 11.84 2
21510 11.84 3
48467 11.88 2
7048 12.00 2
58567 12.07 3
53422 12.14 2

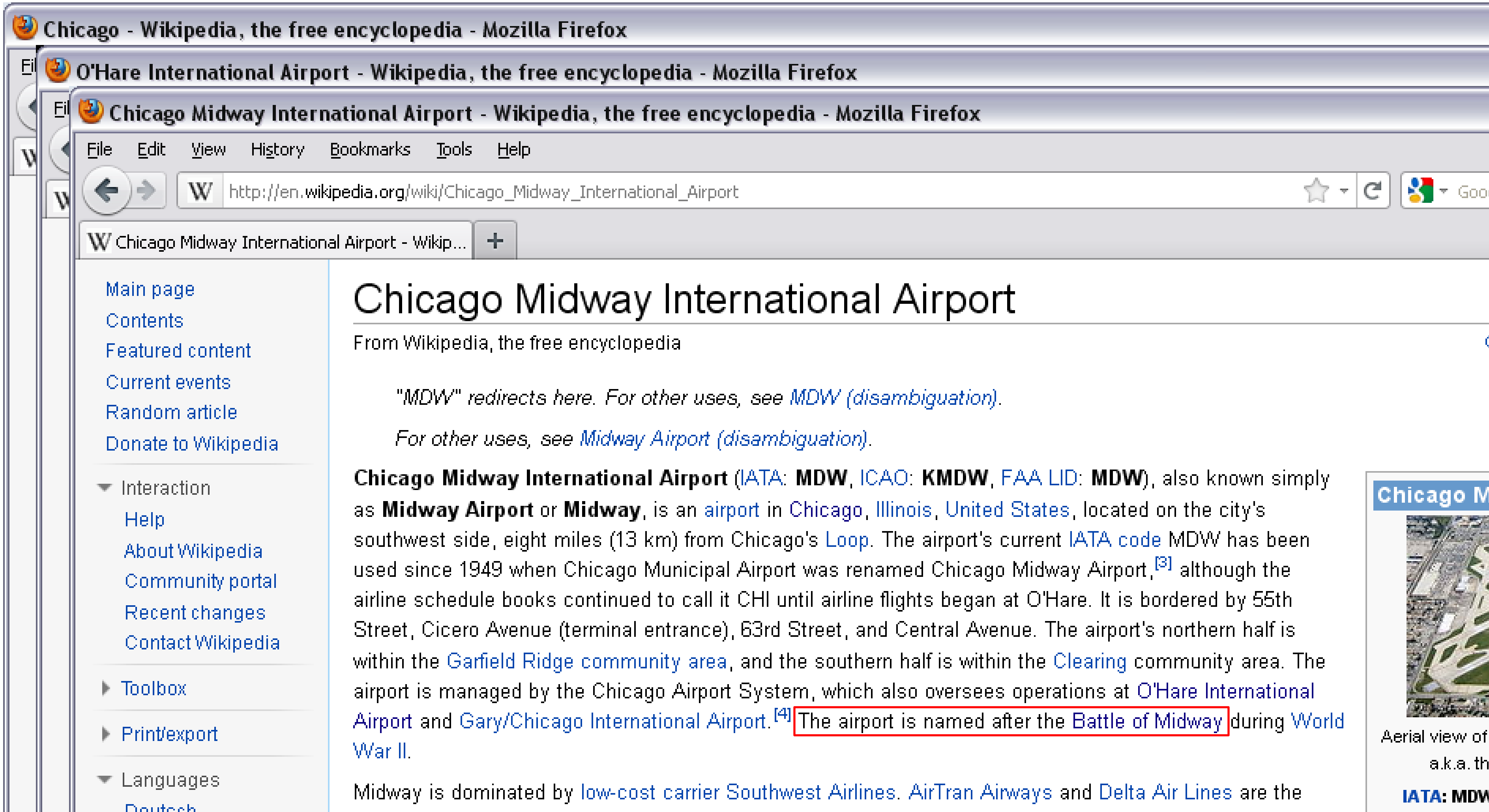
53615 12.16 3
37371 12.22 3
51435 12.26 2
39851 12.30 2
50900 12.53 2
45368 12.57 3
18663 12.58 3
17519 12.62 2

CPU time: 0.17 seconds / Database size: 59895 images

http://wang.ist.psu.edu/IMAGE/

Jeopardy!

A big US city with two airports, one named after a World War II hero, and one named after a World War II battle field?


Chicago - Wikipedia, the free encyclopedia - Mozilla Firefox
O'Hare International Airport - Wikipedia, the free encyclopedia - Mozilla Firefox
Chicago Midway International Airport - Wikipedia, the free encyclopedia - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://en.wikipedia.org/wiki/Chicago_Midway_International_Airport
Chicago Midway International Airport - Wikip...
Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia
Toolbox
Print/export
Languages
Deutsch
Chicago M
Aerial view of
a.k.a. th
IATA: MDW

Chicago Midway International Airport

From Wikipedia, the free encyclopedia

"MDW" redirects here. For other uses, see MDW (disambiguation).

For other uses, see Midway Airport (disambiguation).

Chicago Midway International Airport (IATA: **MDW**, ICAO: **KMDW**, FAA LID: **MDW**), also known simply as **Midway Airport** or **Midway**, is an airport in Chicago, Illinois, United States, located on the city's southwest side, eight miles (13 km) from Chicago's Loop. The airport's current IATA code MDW has been used since 1949 when Chicago Municipal Airport was renamed Chicago Midway Airport,[3] although the airline schedule books continued to call it CHI until airline flights began at O'Hare. It is bordered by 55th Street, Cicero Avenue (terminal entrance), 63rd Street, and Central Avenue. The airport's northern half is within the Garfield Ridge community area, and the southern half is within the Clearing community area. The airport is managed by the Chicago Airport System, which also oversees operations at O'Hare International Airport and Gary/Chicago International Airport.[4] The airport is named after the Battle of Midway during World War II.

Midway is dominated by low-cost carrier Southwest Airlines. AirTran Airways and Delta Air Lines are the

Deep-QA in NL

William Wilkinson's "An Account of the Principalities of Wallachia and Moldavia" inspired this author's most famous novel

This town is known as "Sin City" & its downtown is "Glitter Gulch"

As of 2010, this is the only former Yugoslav republic in the EU

99 cents got me a 4-pack of Ytterlig coasters from this Swedish chain

question classification & decomposition

knowledge backends

D. Ferrucci et al.: Building Watson: An Overview of the DeepQA Project. AI Magazine, 2010.

www.ibm.com/innovation/us/watson/index.htm

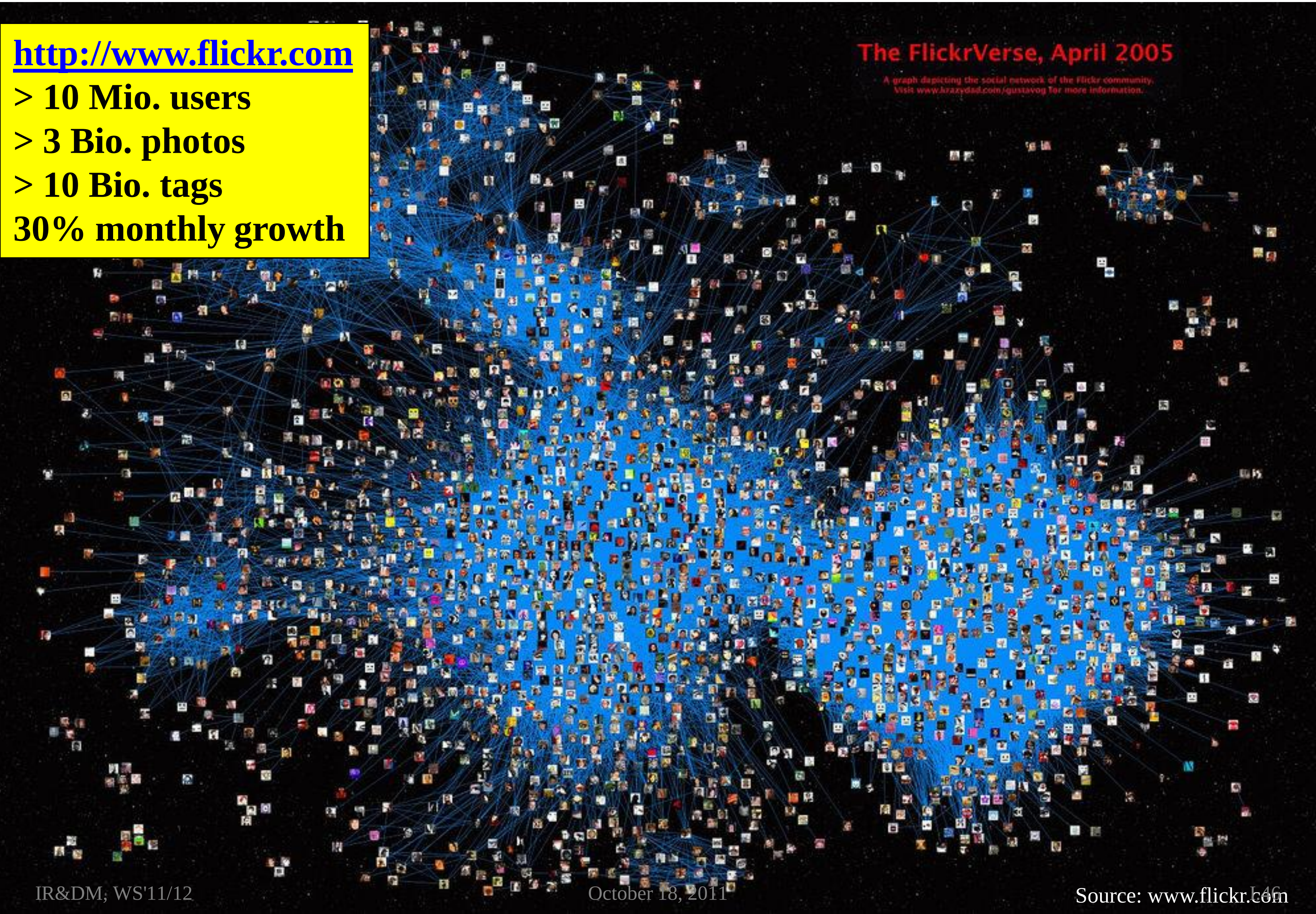
“Wisdom of the Crowds” at Work on Web 2.0

Information enrichment & knowledge extraction **by humans**:

- **Collaborative Recommendations & QA**
  - Amazon (product ratings & reviews, recommended products)
  - Netflix: movie DVD rentals → $ 1 Mio. Challenge
  - answers.yahoo.com, iknow.baidu, www.answers.com, etc.
- **Social Tagging and Folksonomies**
  - del.icio.us: Web bookmarks and tags
  - flickr.com: photo annotation, categorization, rating
  - librarything.com: same for books
- **Human Computing in Game Form**
  - ESP and Google Image Labeler: image tagging
  - labelme.csail.mit.edu: objects in images
  - more games with a purpose at http://www.gwap.com/gwap/
- **Online Communities**
  - dblife.cs.wisc.edu for database research, etc.
  - yahoo! groups, facebook, Google+, studivz, etc. etc.

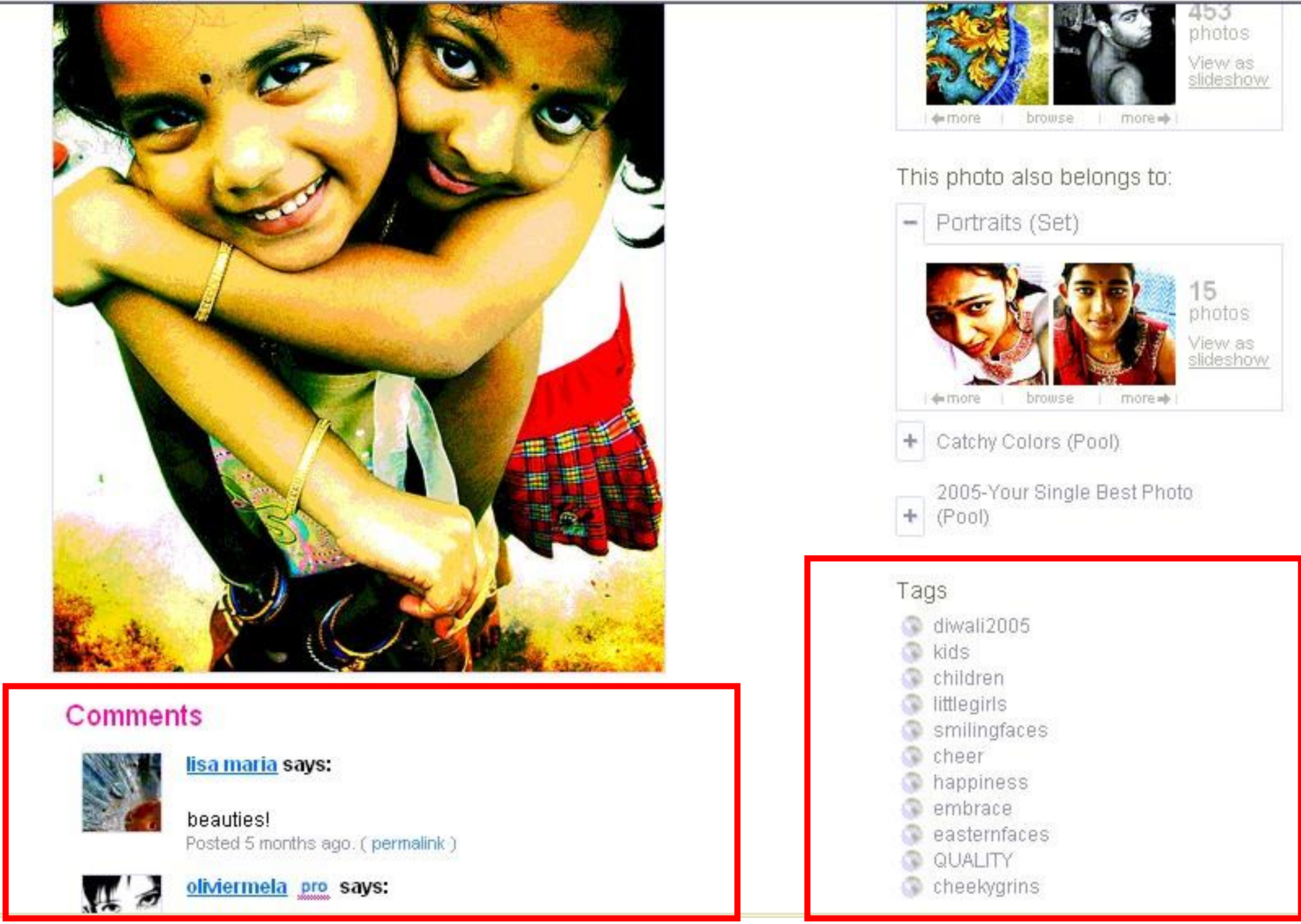
Social-Tagging Community

http://www.flickr.com
> 10 Mio. users
> 3 Bio. photos
> 10 Bio. tags
30% monthly growth


The FlickrVerse, April 2005
A graph depicting the social network of the Flickr community.
Visit www.krazydad.com/gustavog for more information.

Source: www.flickr.com

Social Tagging: Example Flickr



453 photos

View as slideshow

This photo also belongs to:

– Portraits (Set)

15 photos

View as slideshow

+ Catchy Colors (Pool)

+ 2005-Your Single Best Photo (Pool)

Comments

lisa maria says:

beauties!

Posted 5 months ago. (permalink)

oliviermela pro says:

Tags

- diwali2005
- kids
- children
- littlegirls
- smilingfaces
- cheer
- happiness
- embrace
- easternfaces
- QUALITY
- cheekygrins

IRDM Research Literature

Important **conferences** on IR and DM
(see DBLP bibliography for full detail, http://www.informatik.uni-trier.de/~ley/db/)
SIGIR, WSDM, ECIR, CIKM, WWW, KDD, ICDM, ICML, ECML

Important **journals** on IR and DM
(see DBLP bibliography for full detail, http://www.informatik.uni-trier.de/~ley/db/)
TOIS, TOW, InfRetr, JASIST, InternetMath, TKDD, TODS, VLDBJ

Performance **evaluation/benchmarking** initiatives:

- Text Retrieval Conference (TREC), http://trec.nist.gov
- Cross-Language Evaluation Forum (CLEF), www.clef-campaign.org
- Initiative for the Evaluation of XML Retrieval (INEX), http://www.inex.otago.ac.nz/
- KDD Cup, http://www.kdnuggets.com/datasets/kddcup.html & http://www.sigkdd.org/kddcup/index.php