Information Retrieval and Data Mining

Winter Semester 2005/06
Saarland University, Saarbrücken

Prof. Dr. Gerhard Weikum
weikum@mpi-inf.mpg.de
http://www.mpi-inf.mpg.de/departments/d5/teaching/ws05_06/irdm/

Organization

- **Lectures:**
  Tuesday 14-16 and Thursday 14-16 in 45/001
  Office hours Prof. Weikum: appointment by e-mail
- **Assignments / Tutoring Groups**:
  Friday 9-11, 11-13, or 14-16
  Monday 9-11, 11-13, or 13-15
  Paper assignments given out in Tuesday lecture, to be solved until next Tuesday
  First paper assignment given out on Tuesday, Oct 25
  First meetings of tutoring groups: Friday, Nov 4, and Monday, Nov 7

- **Requirements for obtaining 9 credit points:**
  - will be announced in second week

Outline

1. Overview and System Architectures
2. Basics from Probability Theory and Statistics (1)
3. Basics from Probability Theory and Statistics (2)

Part I: Introduction & Foundations

4. Top-k Query Processing and Indexing
5. Advanced IR Models (1)
6. Advanced IR Models (2)
7. Advanced Link Analysis
8. Ontologies and Query Expansion
9. XML Search with Ranking
10. Peer-to-Peer Search

Part II: Information Search

11. Automatic Classification
12. Clustering and Graph Mining
13. Information Extraction (1)
14. Information Extraction (2)
15. Rule Mining

Part III: Information Organization

General Literature (1)

Information Retrieval:

- Soumen Chakrabarti: Mining the Web: Analyis of Hypertext and Semi Structured Data, Morgan Kaufmann, 2002. see also http://www.cse.iitb.ac.in/~soumen/mining-the-web/
- David A. Grossman, Ophir Frieder: Information Retrieval: Algorithms and Heuristics, Springer, 2004.
- Christopher D. Manning, Hinrich Schütze: Foundations of Statistical Natural Language Processing, MIT Press, 1999.
- Ian H. Witten: Managing Gigabytes: Compressing and Indexing Documents and Images, Morgan Kaufmann, 1999.
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto: Modern Information Retrieval, Addison-Wesley, 1999.
- Norbert Fuhr: Information Retrieval, Vorlesung im SS 2005, Universität Duisburg, http://www.is.informatik.uni-duisburg.de/courses/ir_ss05/index.html
- Christopher Manning, Prabhakar Raghavan, Hinrich Schütze: Introduction to Information Retrieval, Cambridge University Press, 2007, http://www-csli.stanford.edu/~schuetze/information-retrieval-book.html ; see also: http://www.stanford.edu/class/cs276/cs276-2005-syllabus.html http://www.stanford.edu/class/cs276a/syllabus2004.html http://www.stanford.edu/class/cs276b/syllabus.html http://www.ims.uni-stuttgart.de/~schuetze/ws2004ir/
- Pierre Baldi, Paolo Frasconi, Padhraic Smyth: Modeling the Internet and the Web - Probabilistic Methods and Algorithms, Wiley & Sons, 2003.
- Max-Planck Institute for Informatics, ADFOCS Summer School 2004, http://www.mpi-inf.mpg.de/conferences/adfocs-04/program.html
- Berkeley School of Information Management and Systems: Search Engines: Technology, Society, and Business, http://www.sims.berkeley.edu/courses/is141/f05/schedule.html

General Literature (2)

Data Mining:

- Jiawei Han, Micheline Kamber: Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.
- Ian H. Witten, Eibe Frank: Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2005. see also http://www.cs.waikato.ac.nz/ml/weka/
- Margaret H. Dunham: Data Mining, Pearson Education, 2003
- David J. Hand, Heikki Mannila, Padhraic Smyth: Principles of Data Mining, MIT Press, 2001.
- Andrew Moore: Statistical Data Mining Tutorials, CMU, http://www.autonlab.org/tutorials/
- Tobias Scheffer, Steffen Bickel: Maschinelles Lernen und Data Mining, Vorlesung SS 2004, Humboldt-Universität, Berlin, http://www.informatik.hu-berlin.de/Forschung_Lehre/wm/index_e.html

Foundations from Statistical Machine Learning:

- Richard O. Duda, Peter E. Hart, David G. Stork: Pattern Classification, Wiley&Sons, 2000.
- Trevor Hastie, Robert Tibshirani, Jerome H. Friedman: Elements of Statistical Learning, Springer, 2001.
- Tom M. Mitchell: Machine Learning, McGraw-Hill, 1997.

General Literature (3)

Foundations from Stochastics:

- **Larry Wasserman: All of Statistics, Springer, 2004.**
  http://www.stat.cmu.edu/~larry/all-of-statistics/index.html
- **George Casella, Roger L. Berger: Statistical Inference, Duxbury, 2002.**
  http://www.stat.ufl.edu/~casella/
- **Arnold Allen: Probability, Statistics, and Queueing Theory with Computer Science Applications, Academic Press, 1990.**

Practical Tools and Programming:

- **Ian H. Witten, Eibe Frank: Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2005.**
  see also http://www.cs.waikato.ac.nz/ml/weka/
- **Erik Hatcher, Otis Gospodnetic: Lucene in Action, Manning Publications, 2004.**
  see also http://lucene.apache.org/
- **Tony Loton: Web Content Mining with Java, John Wiley & Sons, 2002.**

Chapter 1: IRDM Applications and System Architectures

1.1 Overview of IRDM Technologies and Applications
1.2 Web Search Engines
1.3 Towards Semantic Search Engines
1.4 Deep Web Search
1.5 Intranet and Enterprise Search
1.6 Personalized Search and Personal Info Management
1.7 Peer-to-Peer Search and Collaboration
1.8 Multimedia and NLP Search

1.1 Overview of IRDM Applications and 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
- intranet & enterprise search
- XML & text integration
- personalized filtering
- P2P search & collaboration
- multimedia search

Data Mining (DM):

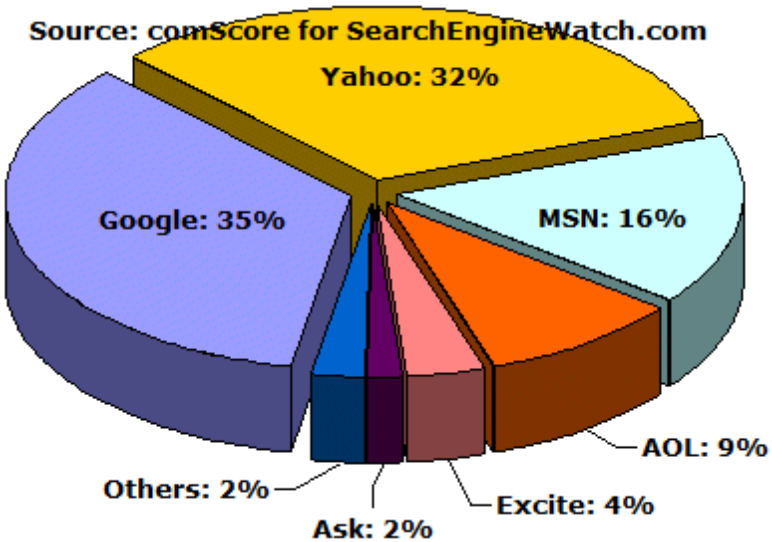
- learning predictive models from data
- pattern, rule, trend, outlier detection
- classification, grouping, segmentation
- knowledge discovery in data collections
- information extraction from text & Web
- graph mining (e.g. on Web graph)

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 & ontology construction

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

1.2 Web Search Engines

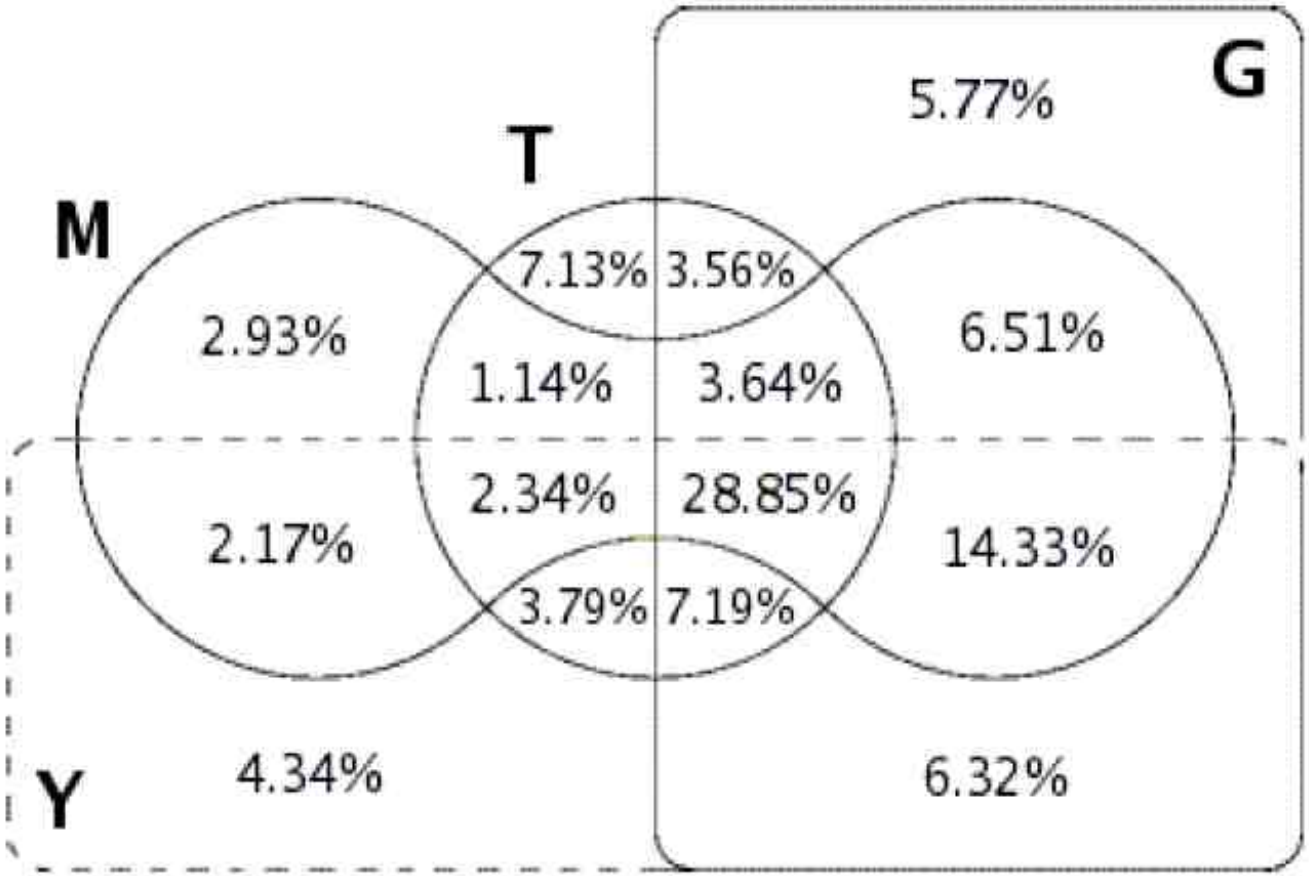


> 11 Billion pages ($11*10^9$)
> 450 Million daily queries
> 8 Billion US $ annual revenue

Outline:

- Web IR basics
- System architecture
- Usage patterns & quality assessment
- Limitations

Web Size and Web Coverage



Source:
A. Gulli, A. Signorini, WWW 2005

Google > 8 Bio., MSN > 5 Bio., Yahoo! > 4 Bio., Ask/Teoma > 2 Bio.
overlap statistics → ***(surface) Web > 11.5 Bio. pages (> 40 TBytes)***

Deep Web (Hidden Web) estimated to have ***500 Bio. units (> 10 PBytes)***

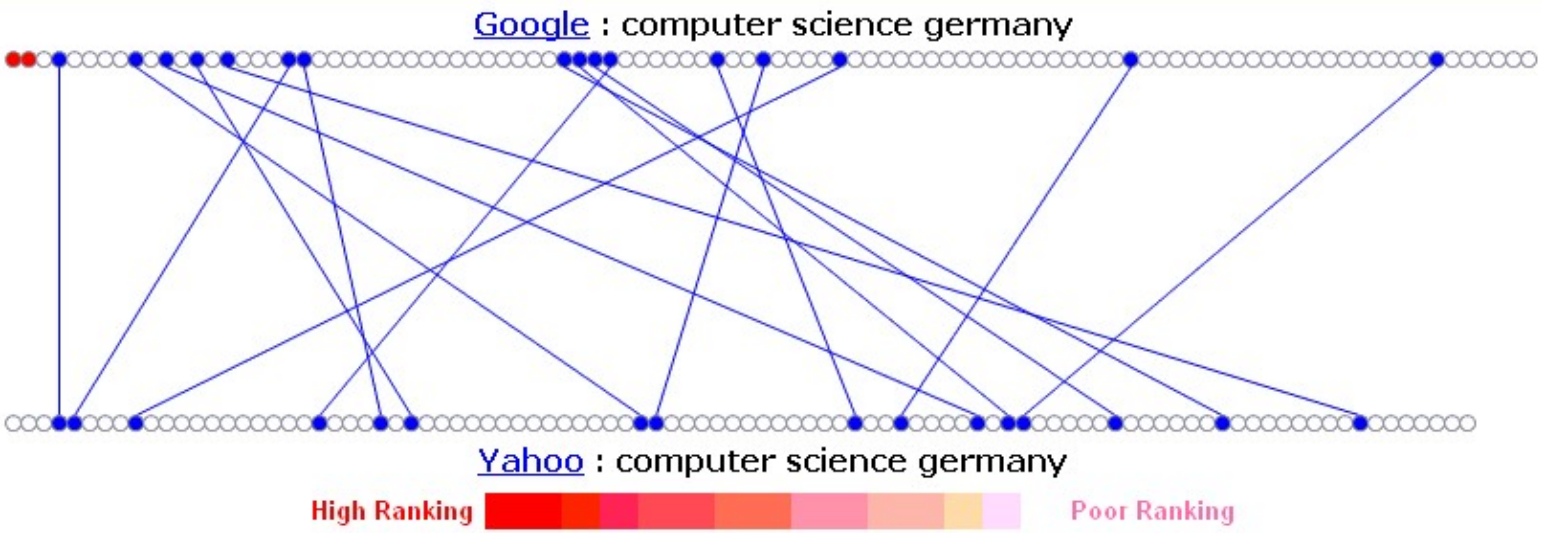
Web Coverage by Search Engines

THUMBSHOTS by Smartdevil RANKING

Search: computer science german | google | versus | computer science german | yahoo | Rank

Highlight Site: www.mpi-sb.mpg.de (optional eg. mysite.com) This service requires IE5+

Due to high traffic, results may not appear. Please try again later.
Support Thumbshots Ranking by donating or sponsoring on Thumbshots.org.



Google (computer science germany)		Yahoo (computer science germany)	
Overlapping Links:	16 (16 %)	Overlapping Links:	16 (17 %)
Unique Links:	84 (84 %)	Unique Links:	80 (83 %)
Total Links:	100	Total Links:	96

http://ranking.thumbshots.com/
http://rankcomparison.di.unipi.it/

Web Archiving

INTERNET ARCHIVE
WayBackMachine

Enter Web Address: http:// All Take Me Back Adv. Search Compare Archive Pages

Searched for http://www.mpi-sb.mpg.de 373 Results

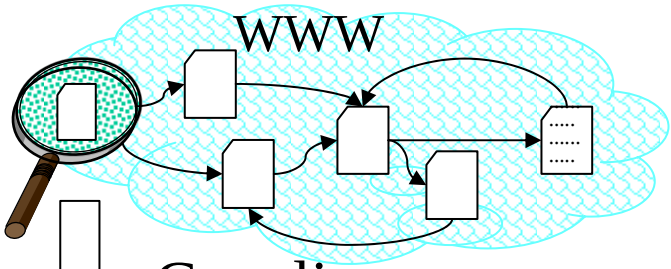
Note some duplicates are not shown. See all.
* denotes when site was updated.

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
http://www.archive.org

Web Content Gathering and Indexing



Crawling

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

Documents

Extraction of relevant words

Surfing
Internet
Cafes
...

Linguistic methods: stemming

Surf
Internet
Cafe
...

Statistically weighted features (terms)

Surf
Wave
Internet
WWW
eService
Cafe
Bistro
...

Indexing

Thesaurus (Ontology)

Synonyms, Sub-/Super-Concepts

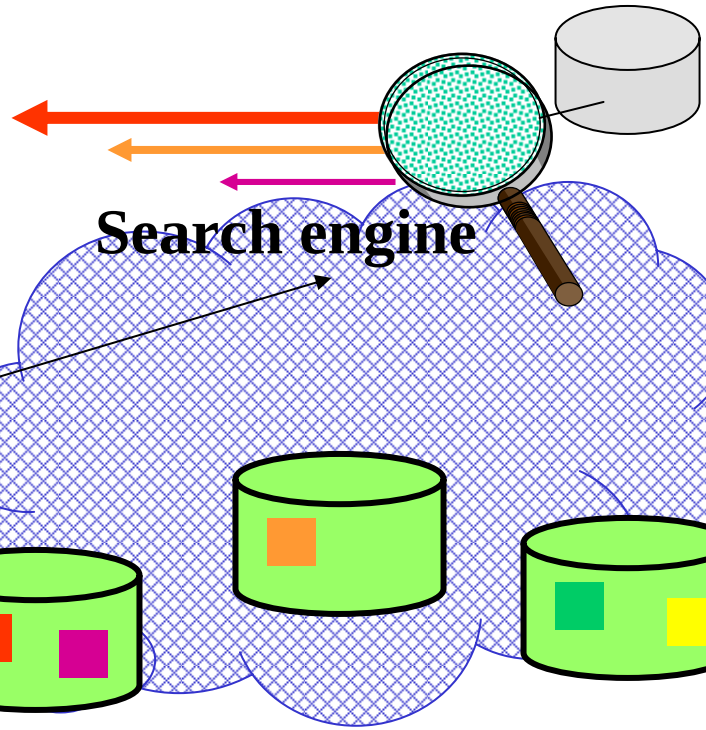
Index (B$^+$-tree)

Bistro | | | Cafe | | | | ...

URLs

Vector Space Model for Content Relevance Ranking

Ranking by descending relevance



Similarity metric:

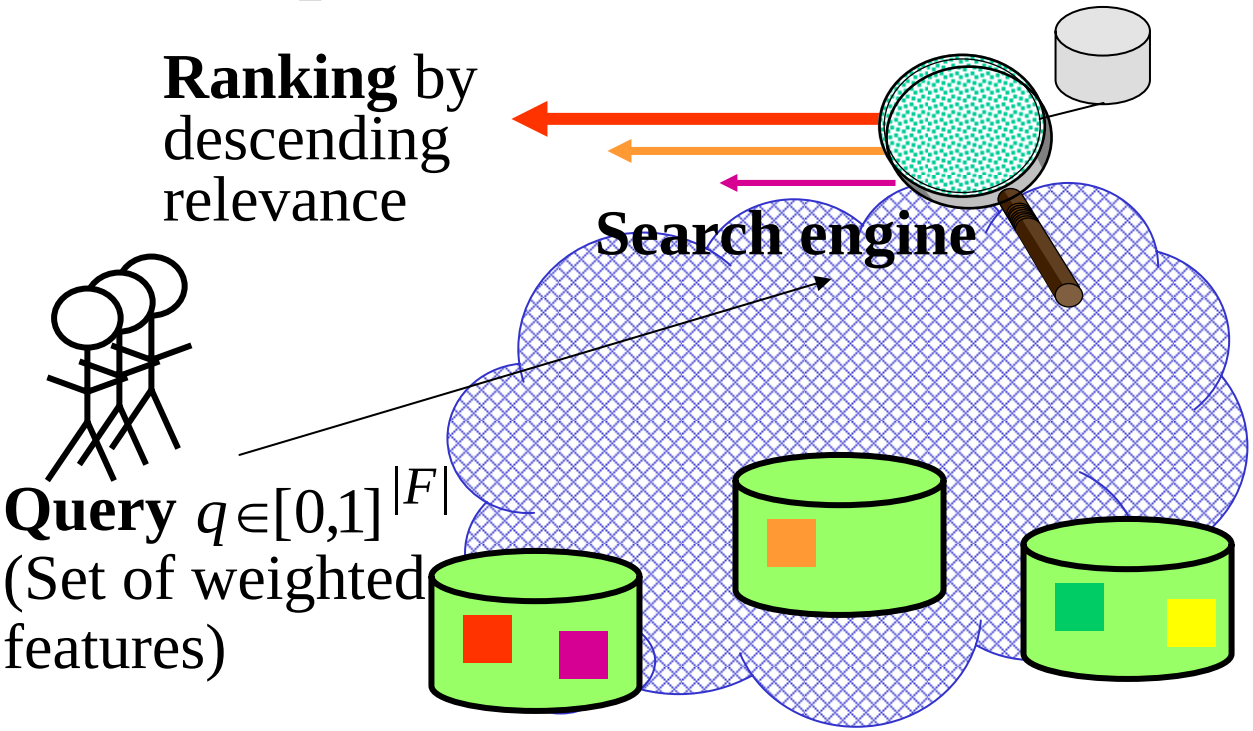
$$sim(d_i, q) := \frac{\sum_{j=1}^{|F|} d_{ij}\, q_j}{\sqrt{\sum_{j=1}^{|F|} d_{ij}^2 \sum_{j=1}^{|F|} q_j^2}}$$

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

Documents are **feature vectors** $d_i \in [0,1]^{|F|}$

Vector Space Model for Content Relevance Ranking

Ranking by descending relevance



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

Similarity metric:

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

Documents are **feature vectors** $d_i \in [0,1]^{|F|}$

e.g., using:

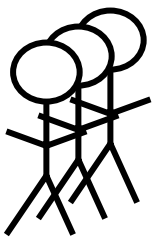
$$d_{ij} := w_{ij} / \sqrt{\Sigma_k w_{ik}^2}$$

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

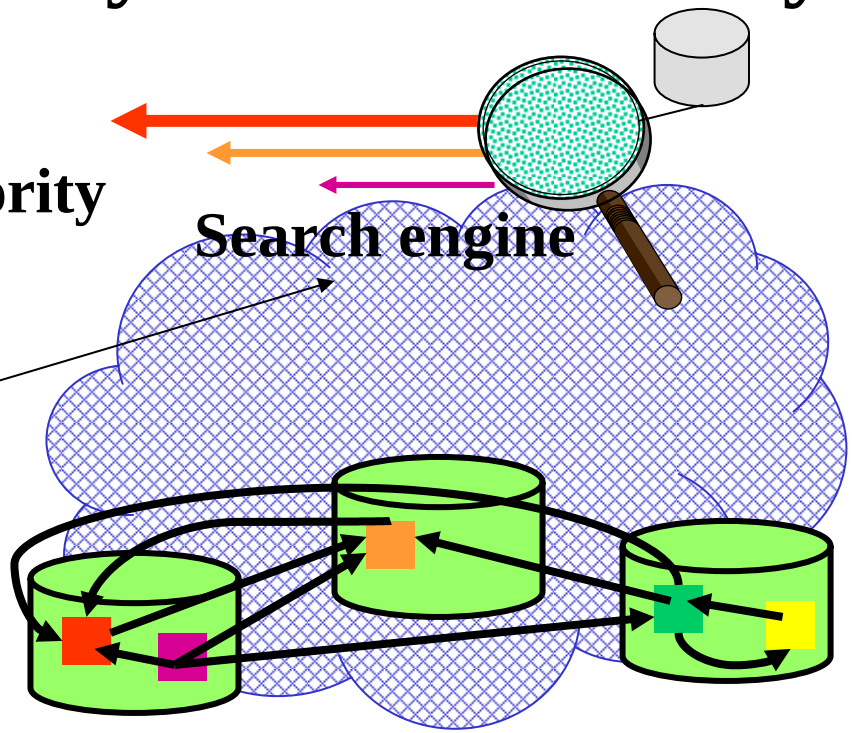
tf*idf formula

Link Analysis for Authority Ranking

Ranking by descending **relevance & authority**



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



+ Consider in-degree and out-degree of Web nodes:

Authority Rank (d_i) :=

Stationary visit probability [d_i]

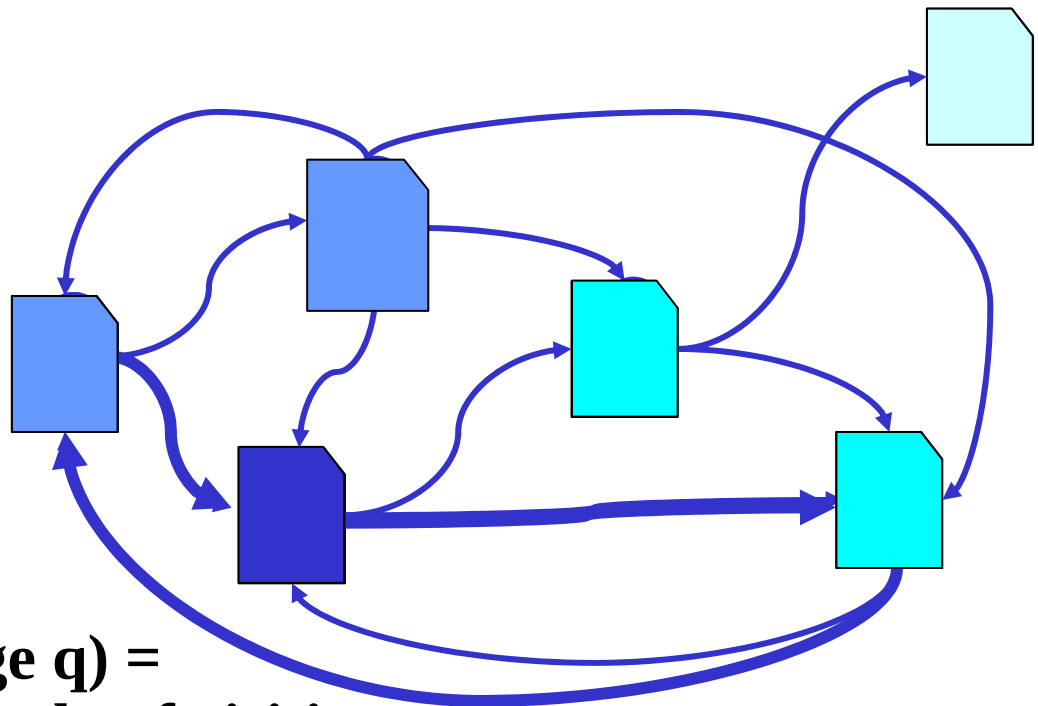
in random walk on the Web

Reconciliation of relevance and authority by ad hoc weighting

Google's PageRank in a Nutshell

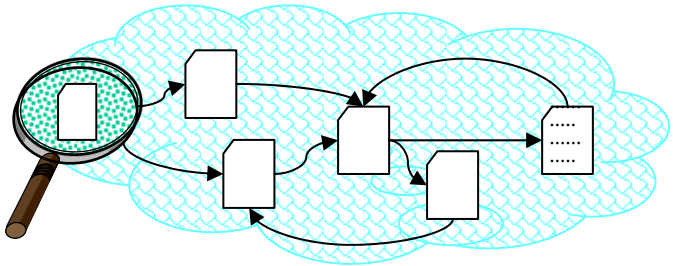
random walk on the Web graph:
uniformly random choice of links + random jumps

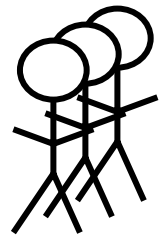
$$PR(q) = \varepsilon \cdot j(q) + (1-\varepsilon) \cdot \sum_{p \in IN(q)} PR(p) \cdot t(p,q)$$



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

System Architecture of a Web Search Engine





crawl | extract & clean | index | search | rank | present

handle dynamic pages, detect duplicates, detect spam

fast top-k queries, query logging and 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

special file system for high-performance storage management

index-entry and query-result caching for fast search

server farm with > 10 000 nodes, distributed/replicated data

Search Engine Users

People who can‘t spell!
[Amit Singhal: SIGIR’05 Keynote]

Search Engine Users

Zeitgeist This Month

Popular Celebrities August 2005
1. madonna
2. jessica simpson
3. pamela anderson
4. paris hilton
5. jessica alba

Music-Related Queries August 2005
1. lyrics
2. my chemical romance
3. beyonce
4. mariah carey
5. green day

Popular References August 2005
1. dictionary
2. maps
3. weather
4. white pages
5. yellow pages

Travel-Related Queries August 2005
1. expedia
2. travelocity
3. orbitz
4. southwest airlines
5. american airlines

Google News Queries

Katrina-Related Queries August 2005
1. hurricane katrina
2. new orleans
3. hurricane katrina photos
4. slidell
5. french quarter

Popular Sports Queries August 2005
1. real madrid
2. arsenal
3. cricket
4. nhl
5. nba

Popular Newsmakers August 2005
1. natalee holloway
2. cindy sheehan
3. peter jennings
4. lance armstrong
5. tiger woods

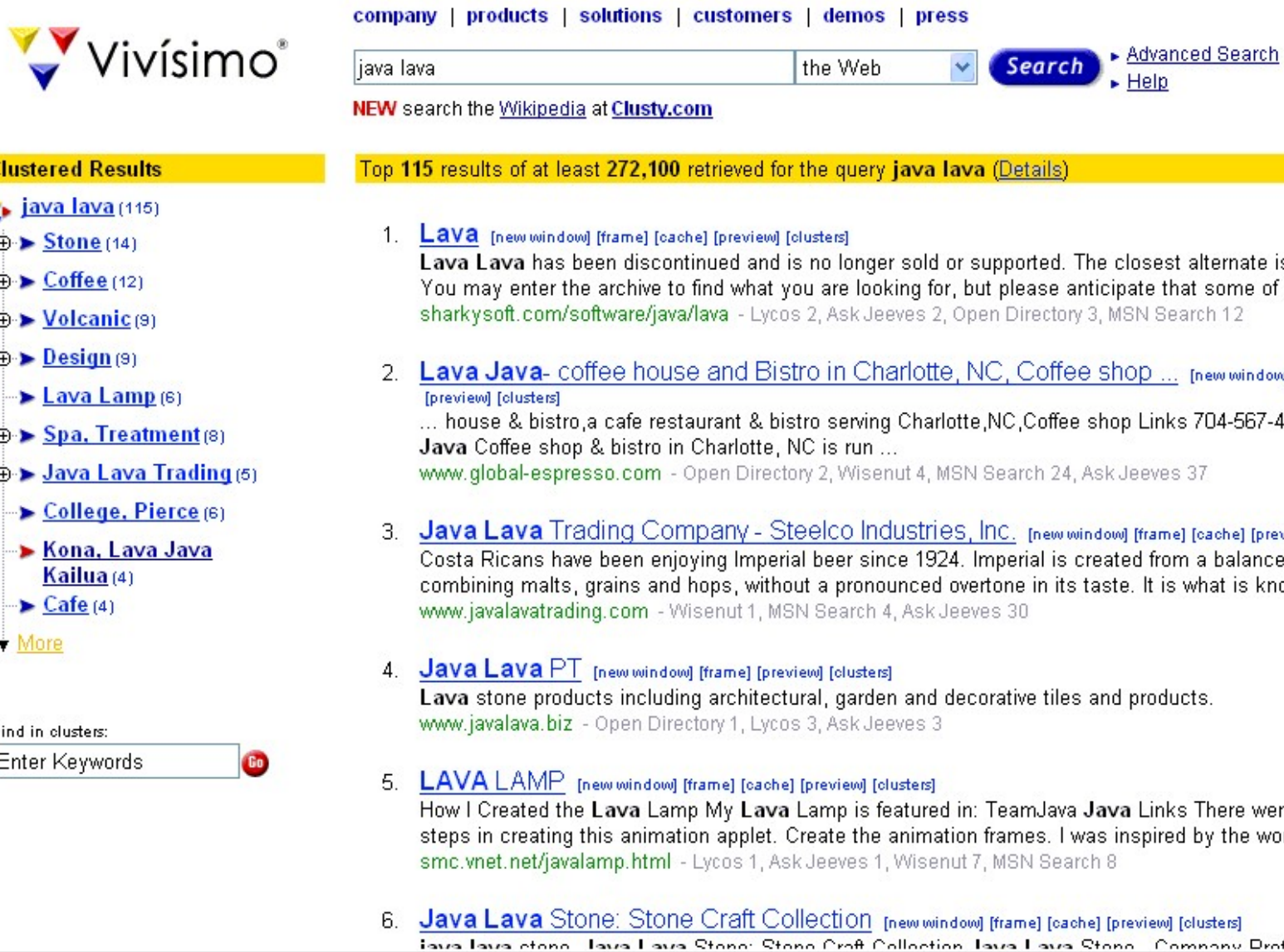
Frequent Web queries:
http://www.google.com/press/zeitgeist.html

Web-Search Usage Patterns

classification of queries [Rose/Levinson: WWW 2004]:

- navigational: find specific homepage with unknown URL, e.g. Cirrus Airlines
- 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
- transactional: find specific resource, e.g. download Lucene source code, Sony Cybershot DSC-W5, Mars surface images, hotel beach south Crete August
- 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

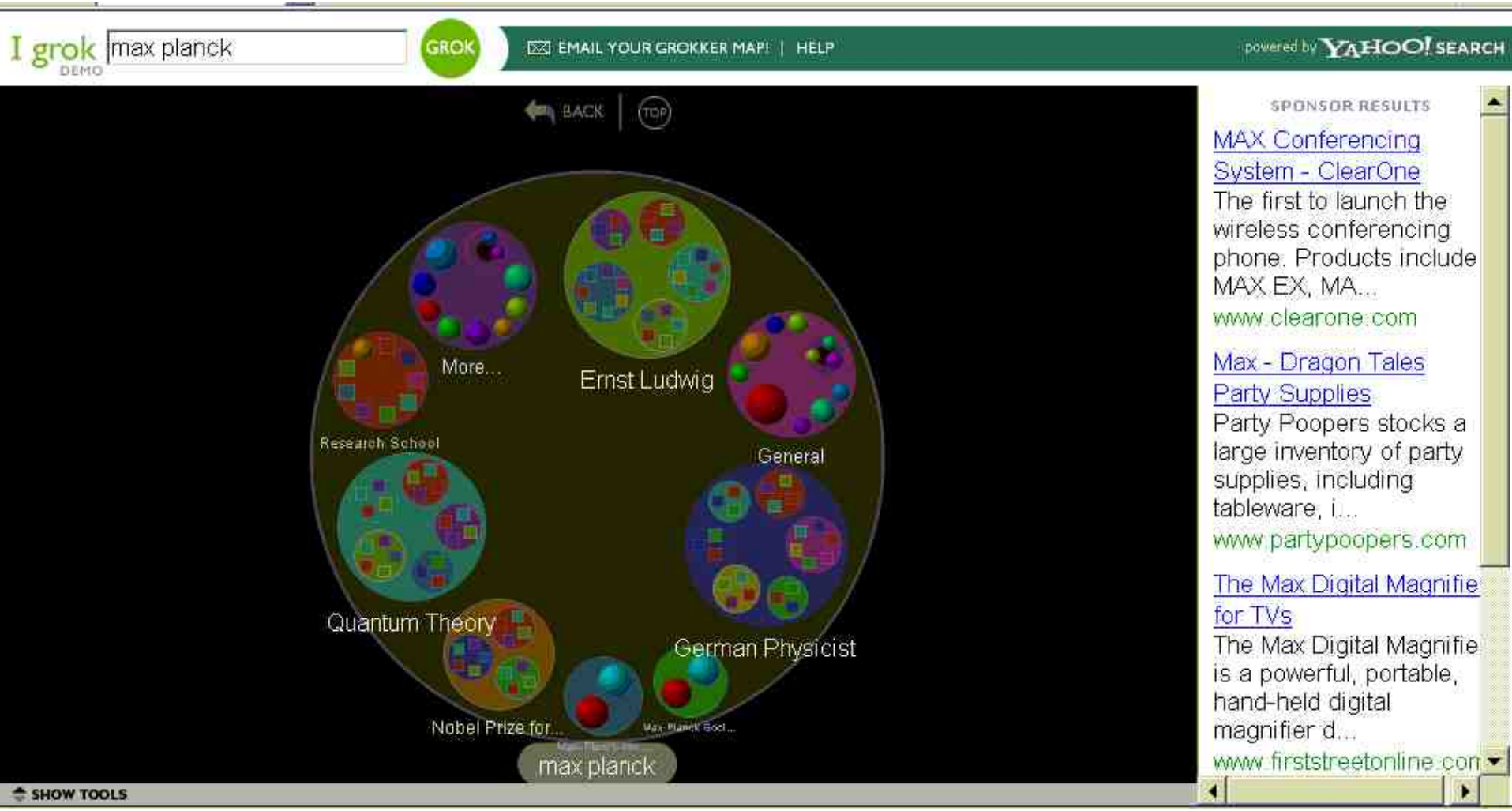
Search Result Organization (1)



cluster search results into topic areas

http://www.vivisimo.com/

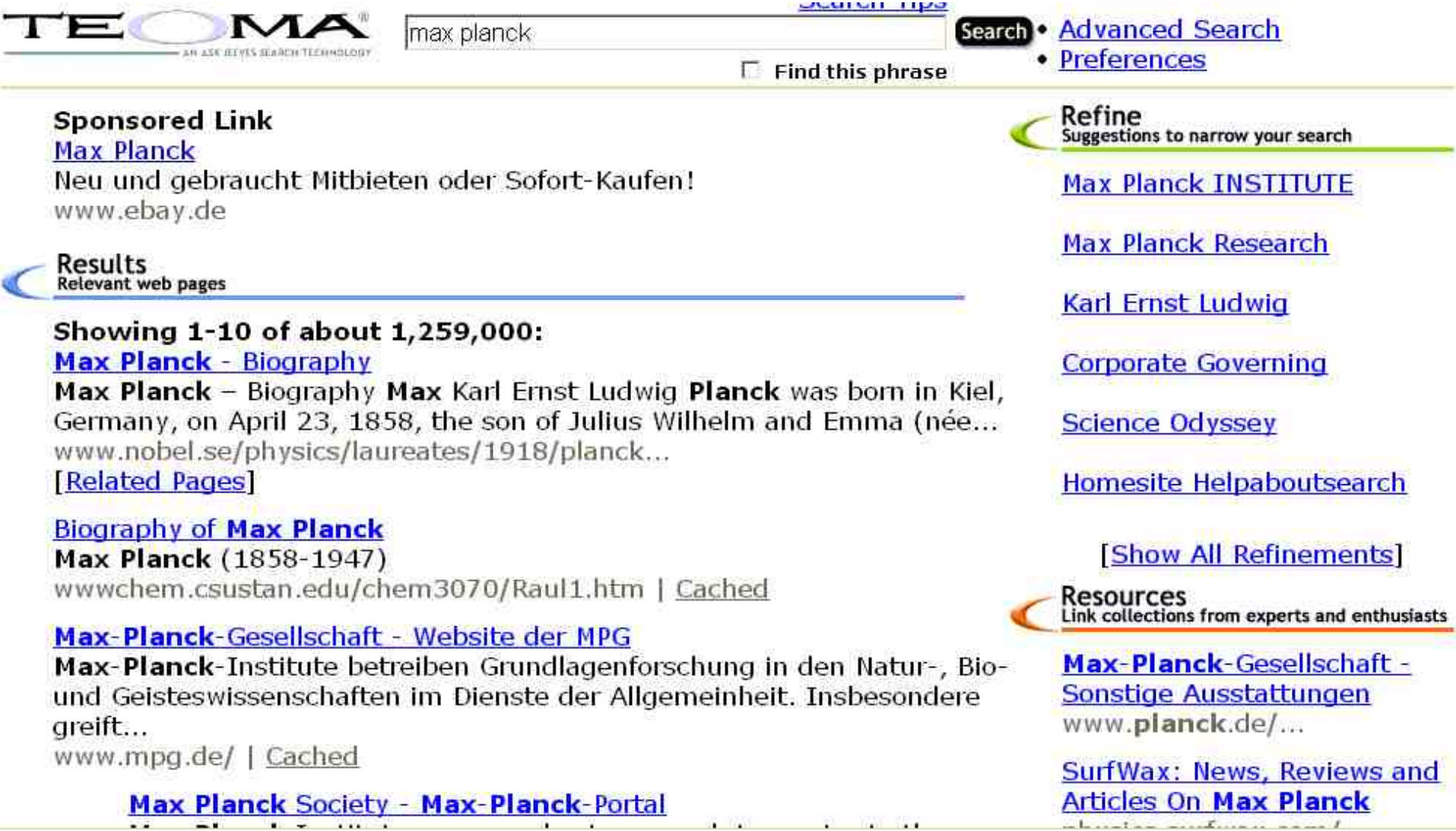
Search Result Organization (3)



visualize cluster hierarchies for search results

http://www.grokker.com/

Search Result Organization (4)

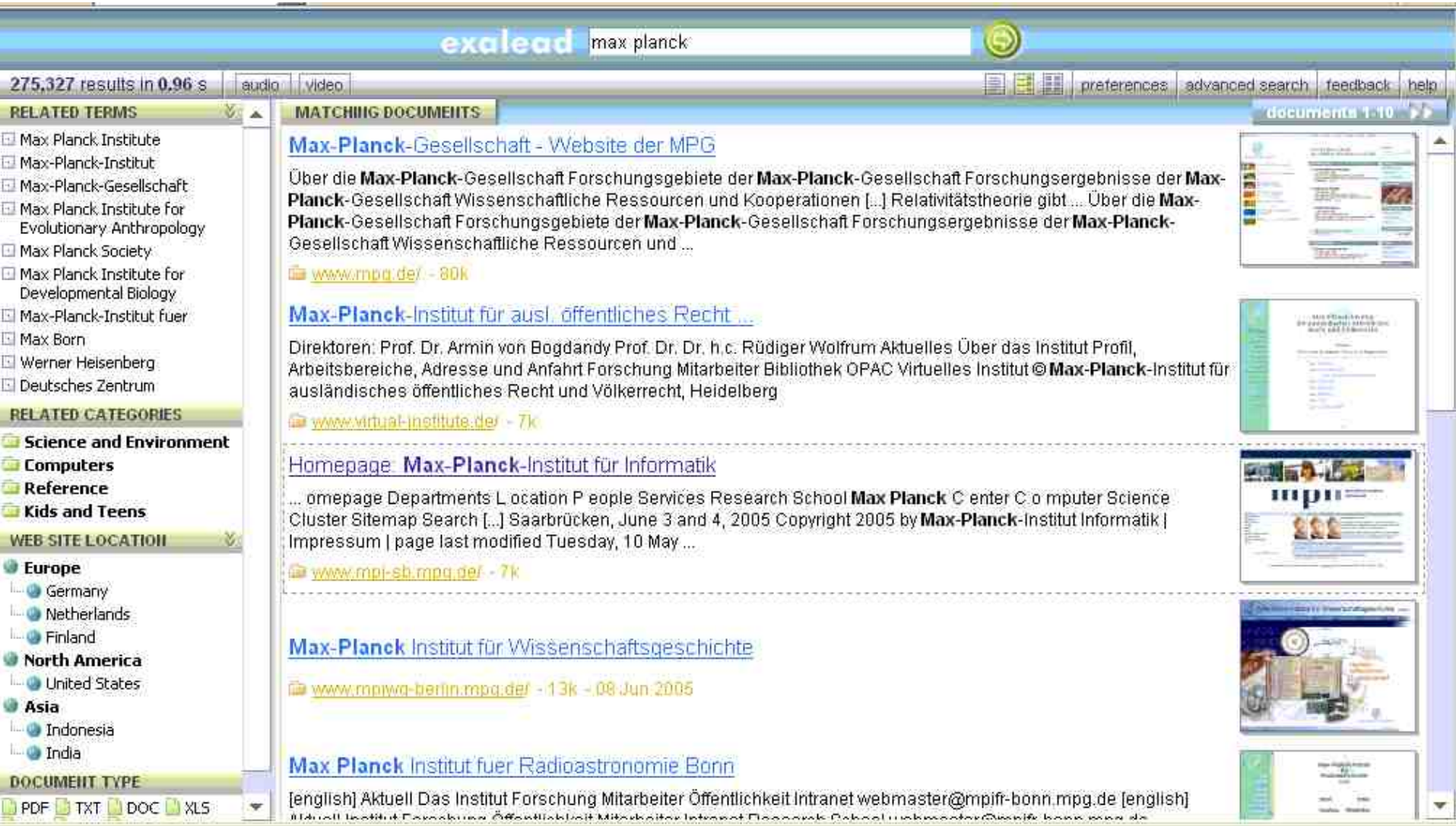


point out related queries
http://www.teoma.com

Search Result Organization (5)

auto-complete queries http://labs.google.com/suggest/
http://www.mpi-inf.mpg.de

Search Result Organization (6)



show broader context of search results

http://www.exalead.com/

Evaluation of Search Result Quality: Basic Measures

ideal measure is user satisfaction
heuristically approximated by benchmarking measures
(on test corpora with query suite and relevance assessment by experts)

Capability to return **only** relevant documents:

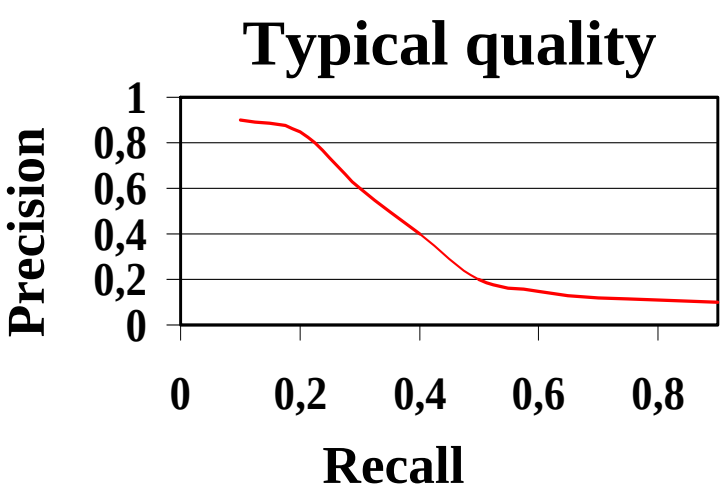
$$\textit{Precision (Präzision)} = \frac{\#\ \textit{relevant docs among top } r}{r}$$

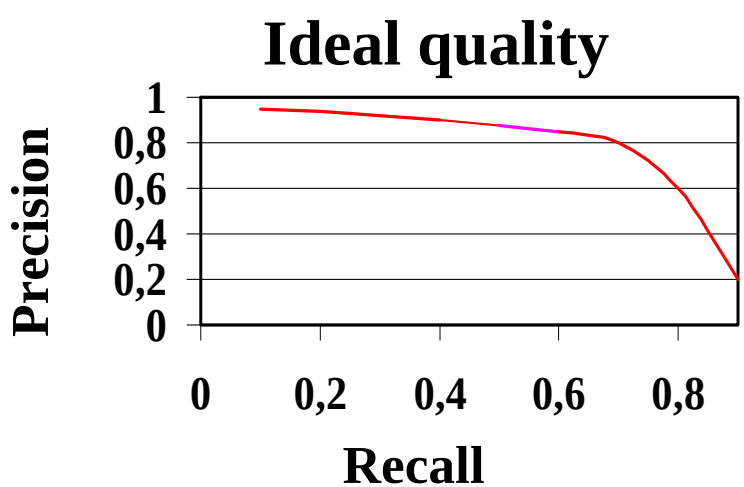
typically for
r = 10, 100, 1000

Capability to return **all** relevant documents:

$$\textit{Recall (Ausbeute)} = \frac{\#\ \textit{relevant docs among top } r}{\#\ \textit{relevant docs}}$$

typically for
r = corpus size





Evaluation of Search Result Quality: Aggregated Measures

Combining precision and recall into **F measure**
(e.g. with α=0.5:
harmonic mean **F1**): $F = \dfrac{1}{\alpha \dfrac{1}{precision} + (1-\alpha)\dfrac{1}{recall}}$

Precision-recall breakeven point of query q:
point on precision-recall curve p = f(r) with p = r

for a set of n queries q1, ..., qn (e.g. TREC benchmark)

Macro evaluation
(user-oriented) of precision $= \dfrac{1}{n}\sum_{i=1}^{n} precision(qi)$

Micro evaluation
(system-oriented) of precision $= \dfrac{\sum_{i=1}^{n} \#\ relevant\ \&\ found\ \ docs\ \ for\ qi}{\sum_{i=1}^{n} \#\ \ found\ \ docs\ \ for\ qi}$

analogous for recall and F1

Evaluation of Search Result Quality: Integrated Measures

- **Interpolated average precision** of query q with precision p(x) at recall x and step width Δ (e.g. 0.1):

$$\frac{1}{1/\Delta} \sum_{i=1}^{1/\Delta} p(i\Delta)$$

area under precision-recall curve

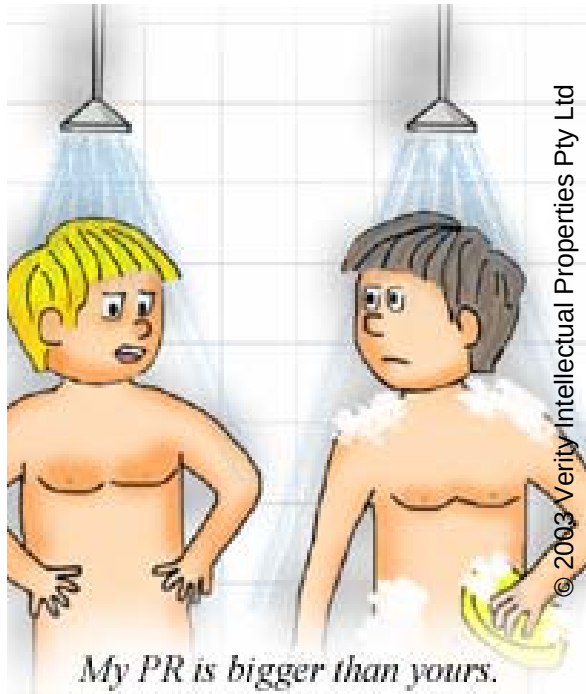
- **Uninterpolated average precision** of query q with top-m search result rank list $d_1, ..., d_m$, relevant results $di_1, ..., di_k$ ($k \leq m$, $i_j \leq i_{j+1} \leq m$):

$$\frac{1}{k} \sum_{j=1}^{k} \frac{j}{i_j}$$

- **Mean average precision (MAP)** of query benchmark suite macro-average of per-query interpolated average precision for top-m results (usually with recall width 0.01)

more measures in the literature

Google & Co: Where Do We Stand Today?



but there are also major limitations!

- great for e-shopping, school kids, scientists, doctors, etc.
- high-precision results for simple queries
- superb scalability (now >8 Bio. docs, >1000 queries/sec)
- continuously enhanced: Froogle, Google Scholar, alerts, multilingual for >100 languages, query auto-completion, etc.

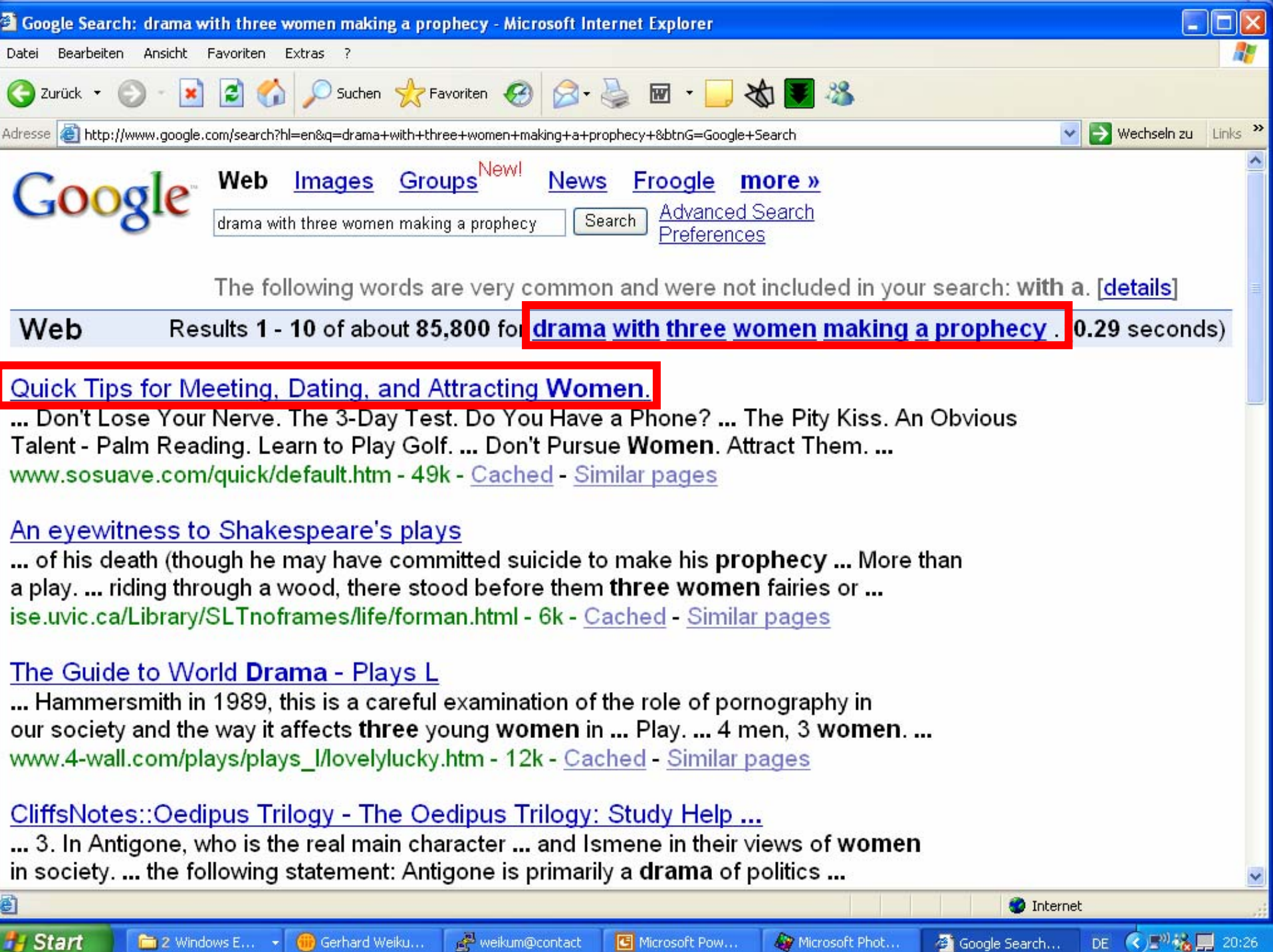
What Google Can‘t Do

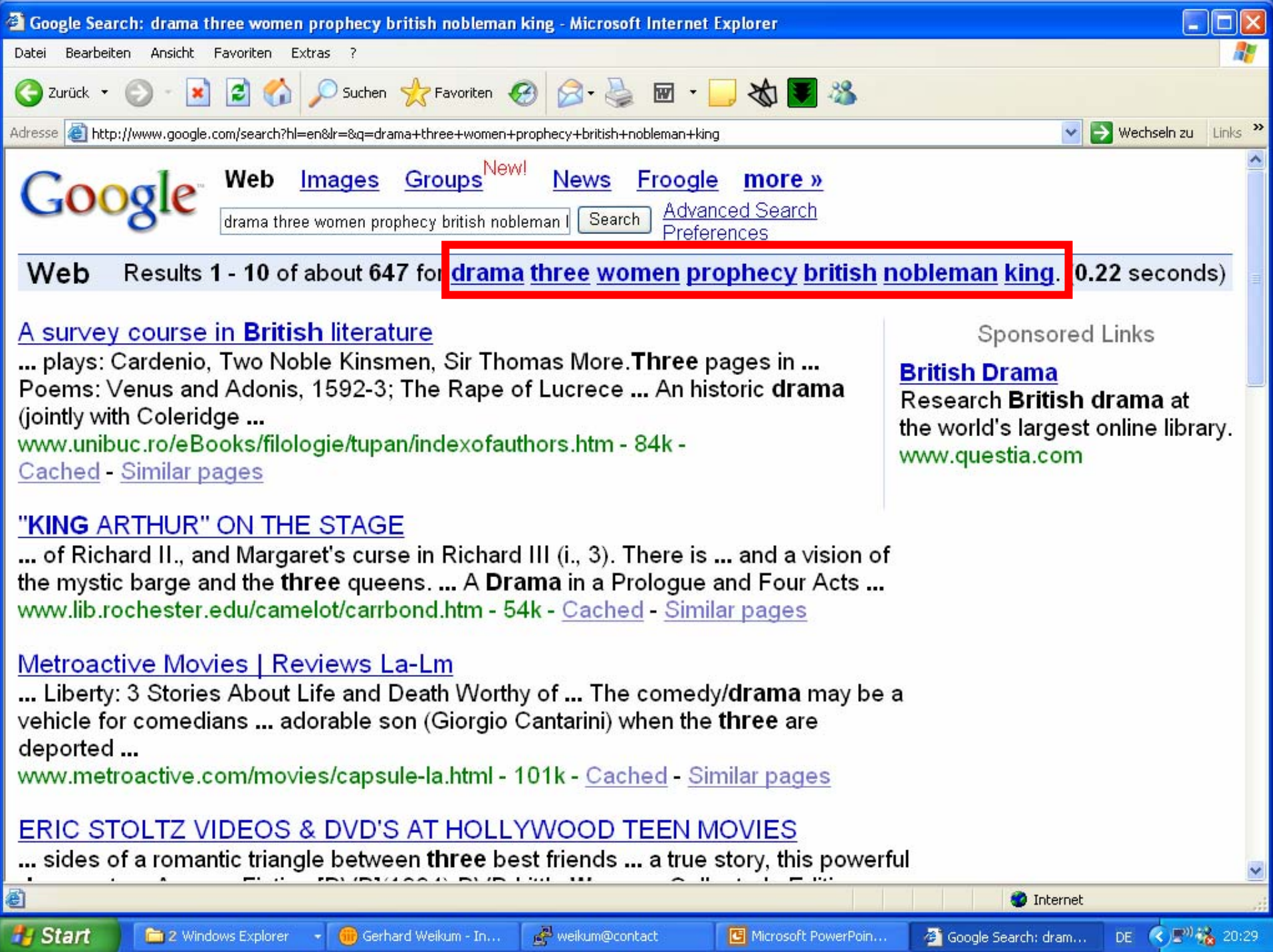
Killer queries (disregarding QA, multilingual, multimedia):

- ***professors from Saarbruecken who teach DB or IR and have projects on XML***
- ***drama with three women making a prophecy to a British nobleman that he will become king***
- ***the woman from Paris whom I met at the PC meeting chaired by Jennifer Widom***

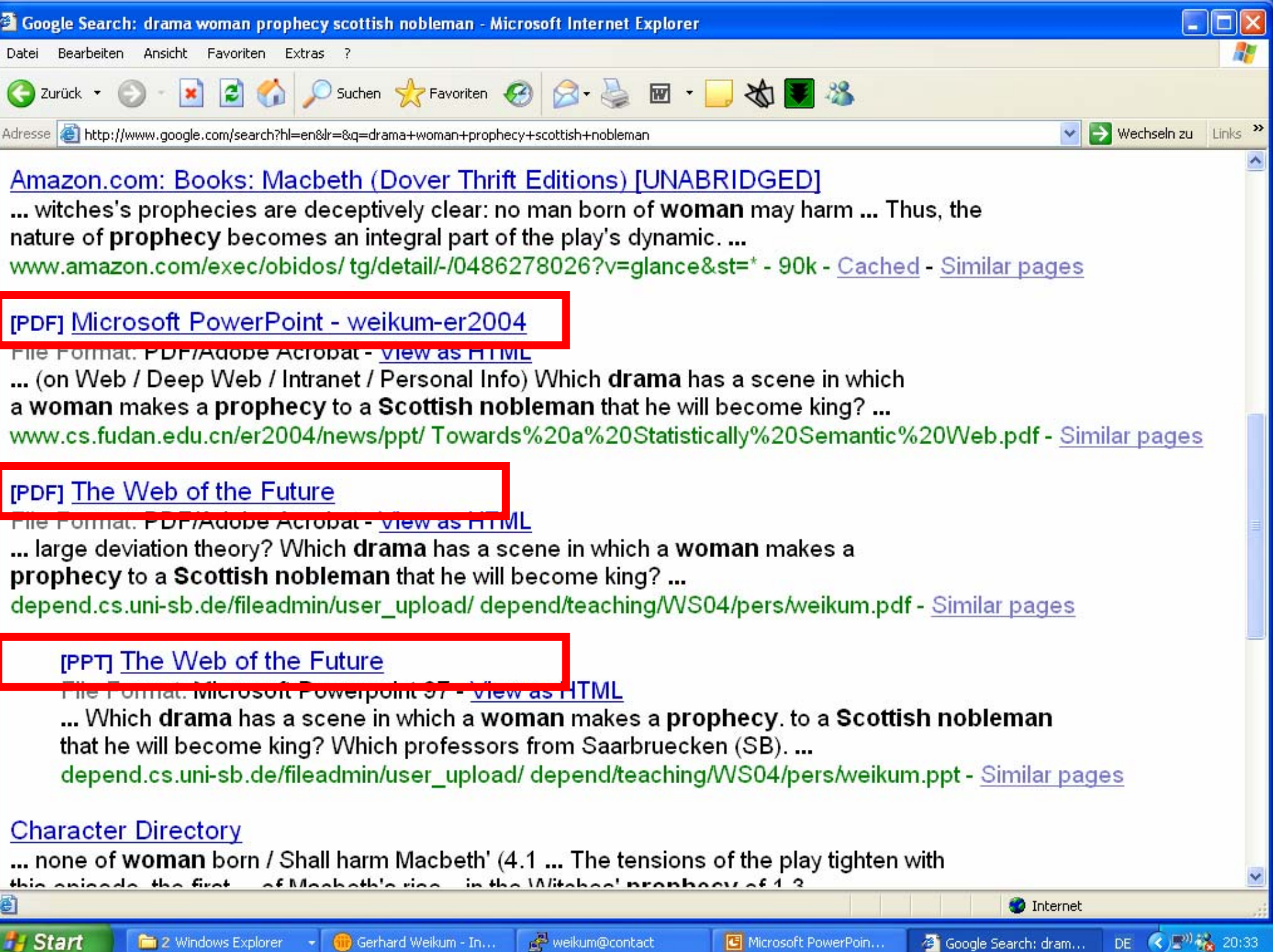
- ***best and latest insights on percolation theory***
- ***pros and cons of dark energy hypothesis***

- ***market impact of XML standards in 2002 vs. 2004***
- ***experienced NLP experts who may be recruited for IT staff***


Google Search: drama with three women making a prophecy - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück Suchen Favoriten
Adresse http://www.google.com/search?hl=en&q=drama+with+three+women+making+a+prophecy+&btnG=Google+Search
Wechseln zu Links
Google
Web Images Groups New! News Froogle more »
drama with three women making a prophecy
Search
Advanced Search
Preferences
The following words are very common and were not included in your search: with a. [details]
Web Results 1 - 10 of about 85,800 for drama with three women making a prophecy . 0.29 seconds)
Quick Tips for Meeting, Dating, and Attracting Women.
... Don't Lose Your Nerve. The 3-Day Test. Do You Have a Phone? ... The Pity Kiss. An Obvious Talent - Palm Reading. Learn to Play Golf. ... Don't Pursue Women. Attract Them. ...
www.sosuave.com/quick/default.htm - 49k - Cached - Similar pages
An eyewitness to Shakespeare's plays
... of his death (though he may have committed suicide to make his prophecy ... More than a play. ... riding through a wood, there stood before them three women fairies or ...
ise.uvic.ca/Library/SLTnoframes/life/forman.html - 6k - Cached - Similar pages
The Guide to World Drama - Plays L
... Hammersmith in 1989, this is a careful examination of the role of pornography in our society and the way it affects three young women in ... Play. ... 4 men, 3 women. ...
www.4-wall.com/plays/plays_l/lovelylucky.htm - 12k - Cached - Similar pages
CliffsNotes::Oedipus Trilogy - The Oedipus Trilogy: Study Help ...
... 3. In Antigone, who is the real main character ... and Ismene in their views of women in society. ... the following statement: Antigone is primarily a drama of politics ...
Internet
Start
2 Windows E...
Gerhard Weiku...
weikum@contact
Microsoft Pow...
Microsoft Phot...
Google Search...
DE
20:26


Google Search: drama three women prophecy british nobleman king - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück Suchen Favoriten
Adresse http://www.google.com/search?hl=en&lr=&q=drama+three+women+prophecy+british+nobleman+king
Wechseln zu Links
Google
Web Images Groups New! News Froogle more »
drama three women prophecy british nobleman k
Search
Advanced Search
Preferences
Web Results 1 - 10 of about 647 for drama three women prophecy british nobleman king. (0.22 seconds)
A survey course in British literature
... plays: Cardenio, Two Noble Kinsmen, Sir Thomas More.Three pages in ... Poems: Venus and Adonis, 1592-3; The Rape of Lucrece ... An historic drama (jointly with Coleridge ...
www.unibuc.ro/eBooks/filologie/tupan/indexofauthors.htm - 84k - Cached - Similar pages
"KING ARTHUR" ON THE STAGE
... of Richard II., and Margaret's curse in Richard III (i., 3). There is ... and a vision of the mystic barge and the three queens. ... A Drama in a Prologue and Four Acts ...
www.lib.rochester.edu/camelot/carrbond.htm - 54k - Cached - Similar pages
Metroactive Movies | Reviews La-Lm
... Liberty: 3 Stories About Life and Death Worthy of ... The comedy/drama may be a vehicle for comedians ... adorable son (Giorgio Cantarini) when the three are deported ...
www.metroactive.com/movies/capsule-la.html - 101k - Cached - Similar pages
ERIC STOLTZ VIDEOS & DVD'S AT HOLLYWOOD TEEN MOVIES
... sides of a romantic triangle between three best friends ... a true story, this powerful
Sponsored Links
British Drama
Research British drama at the world's largest online library.
www.questia.com
Internet
Start
2 Windows Explorer
Gerhard Weikum - In...
weikum@contact
Microsoft PowerPoin...
Google Search: dram...
DE
20:29


Google Search: drama woman prophecy scottish nobleman - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück Suchen Favoriten
Adresse http://www.google.com/search?hl=en&lr=&q=drama+woman+prophecy+scottish+nobleman
Wechseln zu Links
Google
Web Images Groups New! News Froogle more »
drama woman prophecy scottish nobleman
Search
Advanced Search
Preferences
Web Results 1 - 10 of about 456 for drama woman prophecy scottish nobleman (0.25 seconds)
Macbeth
... Lady Macbeth hears of the witches' prophecy, Duncan's ... Lady Macbeth and the three witches are extremely wicked ... Macbeth is the focus of the drama's moral ...
www.sparknotes.com/shakespeare/macbeth/section1.html - 38k - Cached - Similar pages
Macbeth
... to Orlando theater, is a strongly sensual Lady Macbeth, whose ... eyes at the witches' first apparently ridiculous prophecy. ... You see Macbeth as play-actor, the man ...
www.shakespearefest.org/macbeth_99.htm - 27k - Cached - Similar pages
[PDF] Macbeth
File Format: PDF/Adobe Acrobat - View as HTML
... the new titles and appears afraid of the prophecy. ... womb and so "not born of woman." Macbeth conquers ... different treatments in historical source than in drama. ...
www.openstage.com/productions/macbethguid.pdf - Similar pages
[PDF] Macbeth Study Guide 2004.indd
File Format: PDF/Adobe Acrobat - View as HTML
... fate and that they truly prophecy, leaving Macbeth ... son about her husband being gone: Lady Macduff: "How ... speech used by common people in Shakespearean drama. ...
Internet
Start
2 Windows Explorer
Gerhard Weikum - In...
weikum@contact
Microsoft PowerPoin...
Google Search: dram...
DE
20:33


Google Search: drama woman prophecy scottish nobleman - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück Suchen Favoriten
Adresse http://www.google.com/search?hl=en&lr=&q=drama+woman+prophecy+scottish+nobleman
Wechseln zu Links
Amazon.com: Books: Macbeth (Dover Thrift Editions) [UNABRIDGED]
... witches's prophecies are deceptively clear: no man born of woman may harm ... Thus, the nature of prophecy becomes an integral part of the play's dynamic. ...
www.amazon.com/exec/obidos/ tg/detail/-/0486278026?v=glance&st=* - 90k - Cached - Similar pages
[PDF] Microsoft PowerPoint - weikum-er2004
File Format: PDF/Adobe Acrobat - View as HTML
... (on Web / Deep Web / Intranet / Personal Info) Which drama has a scene in which a woman makes a prophecy to a Scottish nobleman that he will become king? ...
www.cs.fudan.edu.cn/er2004/news/ppt/ Towards%20a%20Statistically%20Semantic%20Web.pdf - Similar pages
[PDF] The Web of the Future
File Format: PDF/Adobe Acrobat - View as HTML
... large deviation theory? Which drama has a scene in which a woman makes a prophecy to a Scottish nobleman that he will become king? ...
depend.cs.uni-sb.de/fileadmin/user_upload/ depend/teaching/WS04/pers/weikum.pdf - Similar pages
[PPT] The Web of the Future
File Format: Microsoft Powerpoint 97 - View as HTML
... Which drama has a scene in which a woman makes a prophecy. to a Scottish nobleman that he will become king? Which professors from Saarbruecken (SB). ...
depend.cs.uni-sb.de/fileadmin/user_upload/ depend/teaching/WS04/pers/weikum.ppt - Similar pages
Character Directory
... none of woman born / Shall harm Macbeth' (4.1 ... The tensions of the play tighten with
Internet
Start
2 Windows Explorer
Gerhard Weikum - In...
weikum@contact
Microsoft PowerPoin...
Google Search: dram...
DE
20:33

1.3 Towards Semantic Search Engines

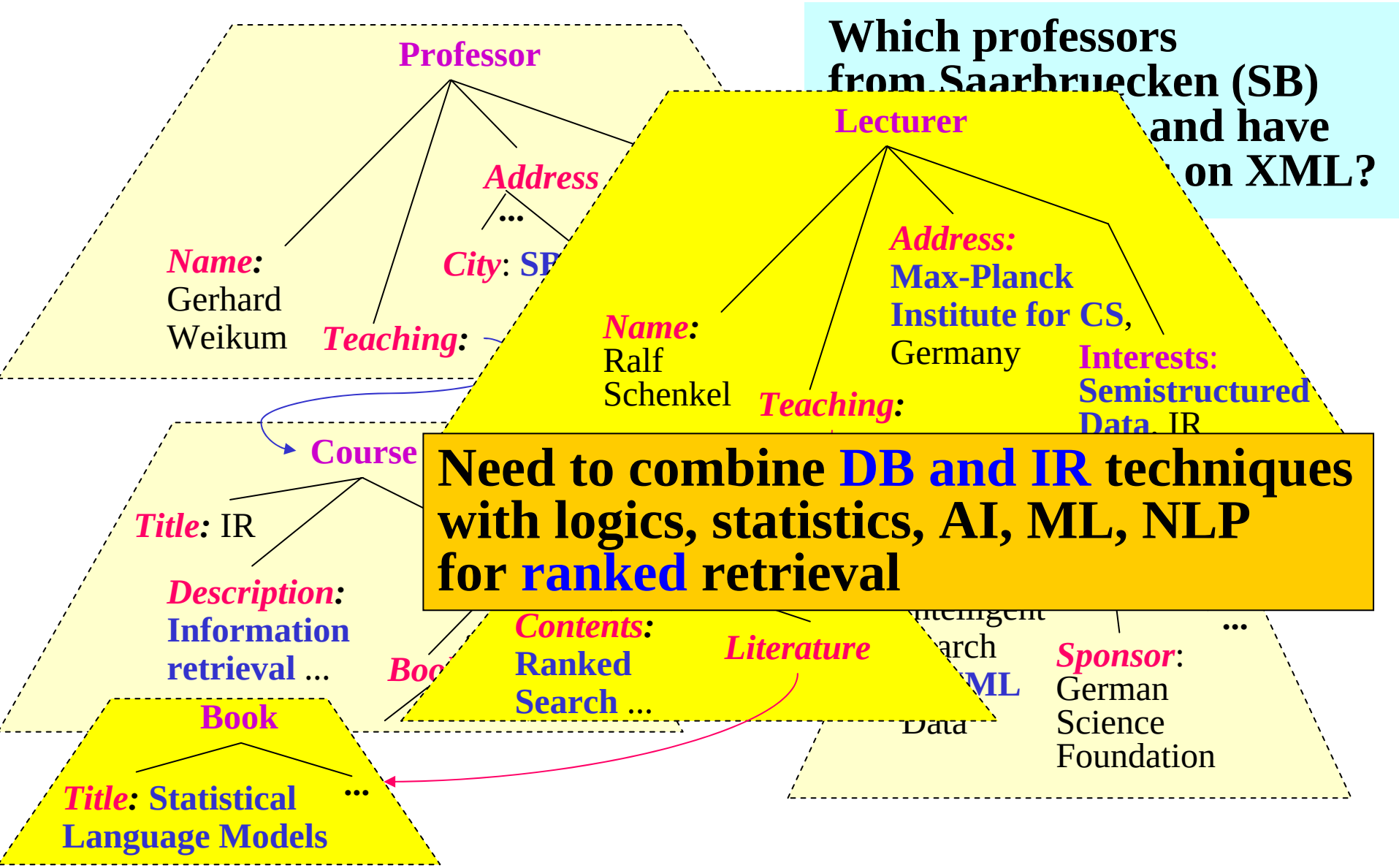
better search result quality (with low human effort) needs:

- **richer content representation** and
- **context awareness** (user profile, place&time, info authors, etc.)

strategic research directions:

- ***background knowledge***
  → ontologies & thesauri, statistics
- ***(semi-)structured and „semantic“ data***
  → metadata, XML, info extraction, annotation & classification
- ***personalization***
  → geo & time, user behavior
- ***humans in the loop***
  → collaboration, recommendation, P2P networks

Rich Content Representation in XML



„Semantic Search“ with TopX Engine

User query: *~c = ~t1 ... ~tm*

Example:

~professor and (~course = „~IR“)

//professor[//place = „SB“]//course = „IR“

Term2Concept with WSD

Query expansion

exp(ti)={w | sim(ti,w)≥ θ}

Weighted expanded query

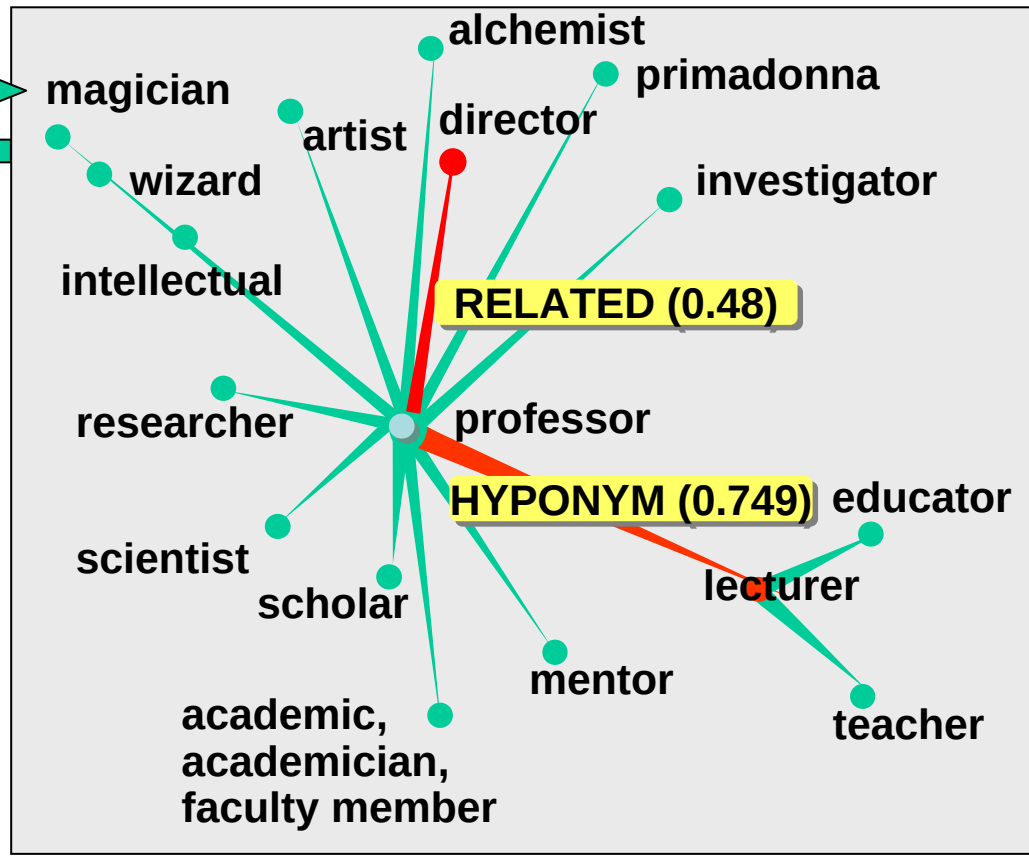
Example:

(professor lecturer (0.749) scholar (0.71) ...)
and ((course class (1.0) seminar (0.84) ...)
= („IR“ „Web search“ (0.653) ...))

Efficient top-k search with dynamic expansion

better recall, better mean precision for hard queries

Thesaurus/Ontology:
concepts, relationships, glosses
from WordNet, Gazetteers,
Web forms & tables, Wikipedia



relationships quantified by statistical correlation measures

Towards a Statistically Semantic Web

Information extraction yields:

(via reg. expr., lexicon, HMM, MRF, etc.)

Person	TimePeriod	...
Sir Isaac Newton ... Leibniz ... Kneller	4 Jan 1643 - ...	

Publication	Topic
Philosophiae Naturalis	... gravitation

Author	Publication
... Newton	Philosophia ...

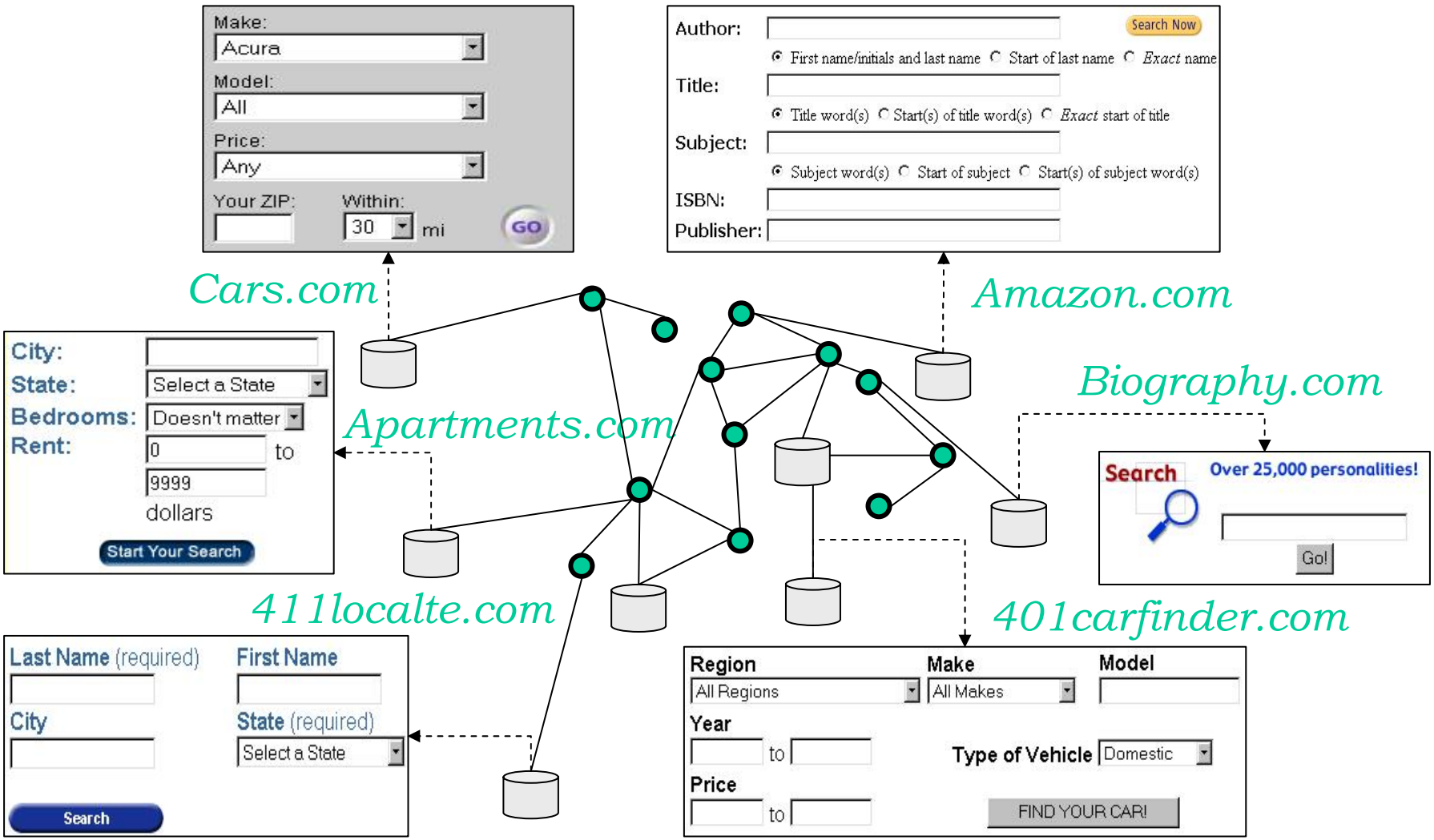
Scientist
Sir Isaac Newton ... Leibniz

but with confidence < 1

→ **Semantic-Web database with uncertainty !**

→ **ranked XML/DB retrieval !**

1.4 Deep Web Search



Source: Kevin Chen-Chuan Chang, CIDR 2005

Deep Web Sources

Data accessible only through query interfaces
(HTML forms, WSDL web services)

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

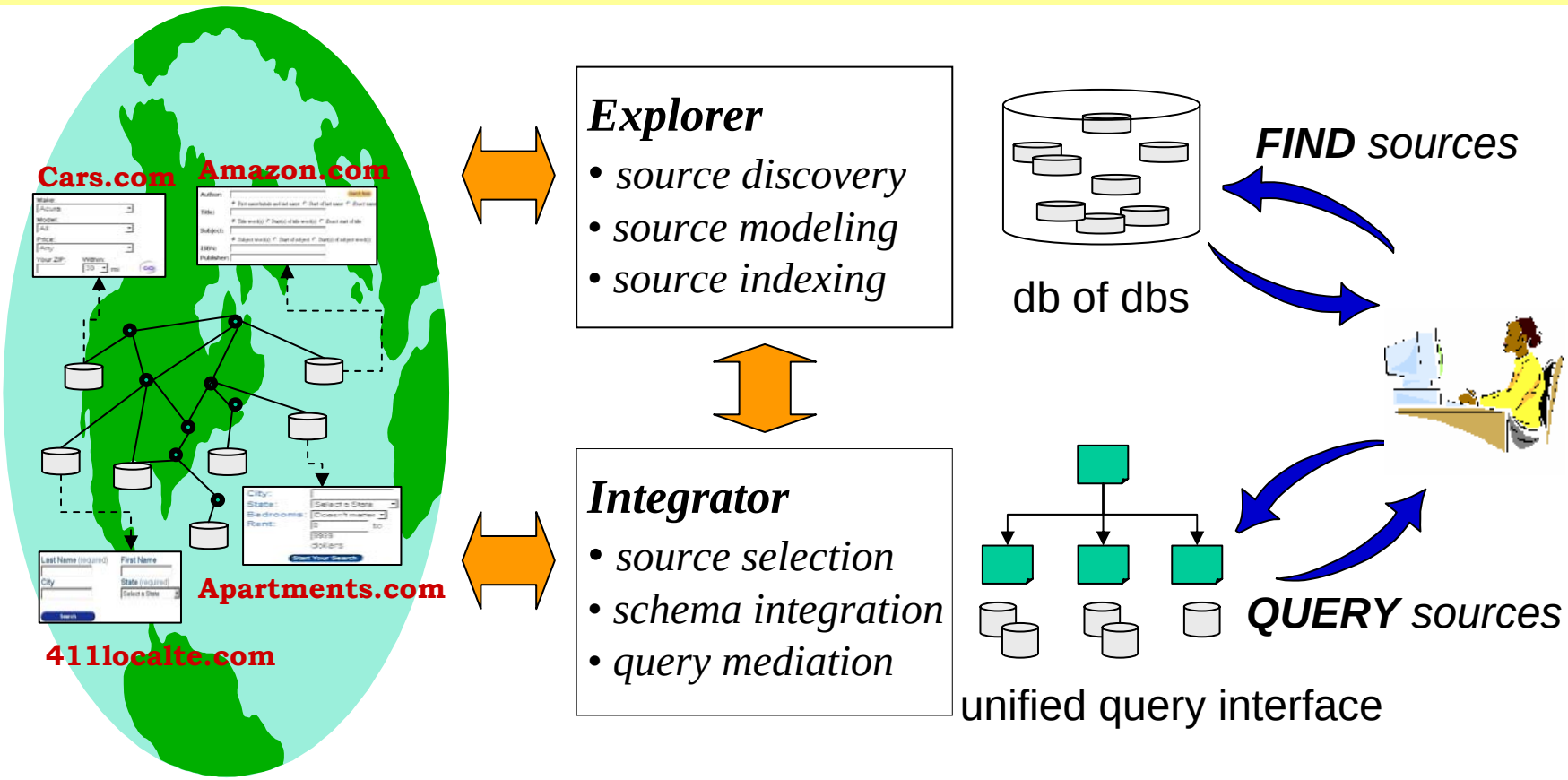
news, libraries, society: cnn.com, yahoo.com, spiegel.de, deutschland.de, uspto.gov, loc.gov, dip.bundestag.de, destatis.de, ddb.de, bnf.fr, kb.nl, kb.se, weatherimages.org, TerraServer.com, lonelyplanet.com

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

Deep Web Research Issues

- find important/relevant sources in the Deep Web (aka. Hidden Web)
- map user queries onto source-specific interfaces

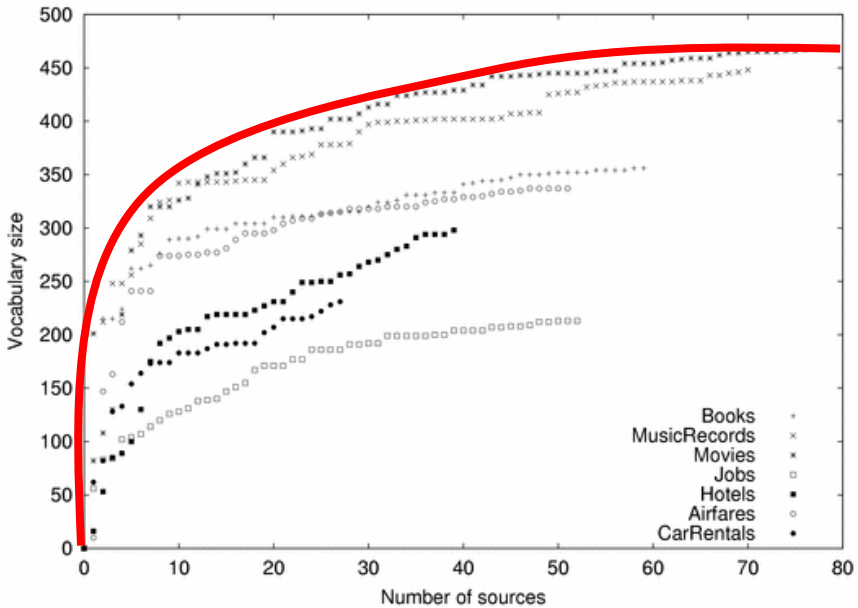
  (metasearch engines are a simple special case)
- merge results into global ranking



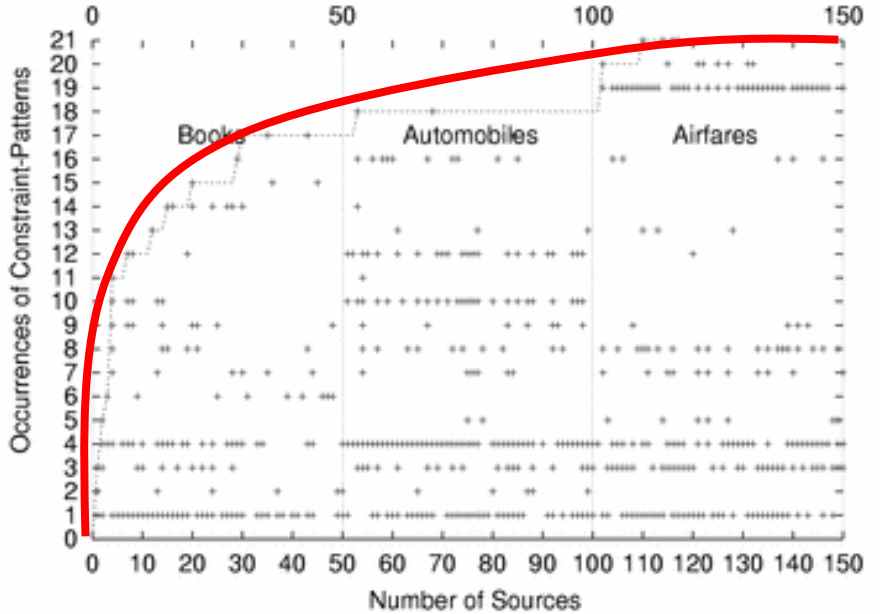
Source: Kevin Chen-Chuan Chang, CIDR 2005

Statistics on Diversity of Names

Attributes converge in a domain!



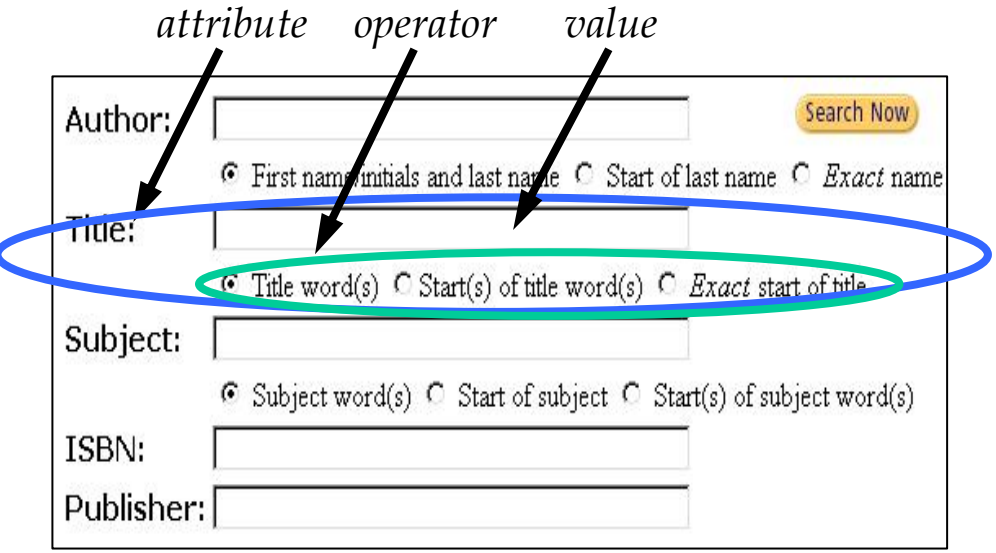
Condition patterns converge even across domains!

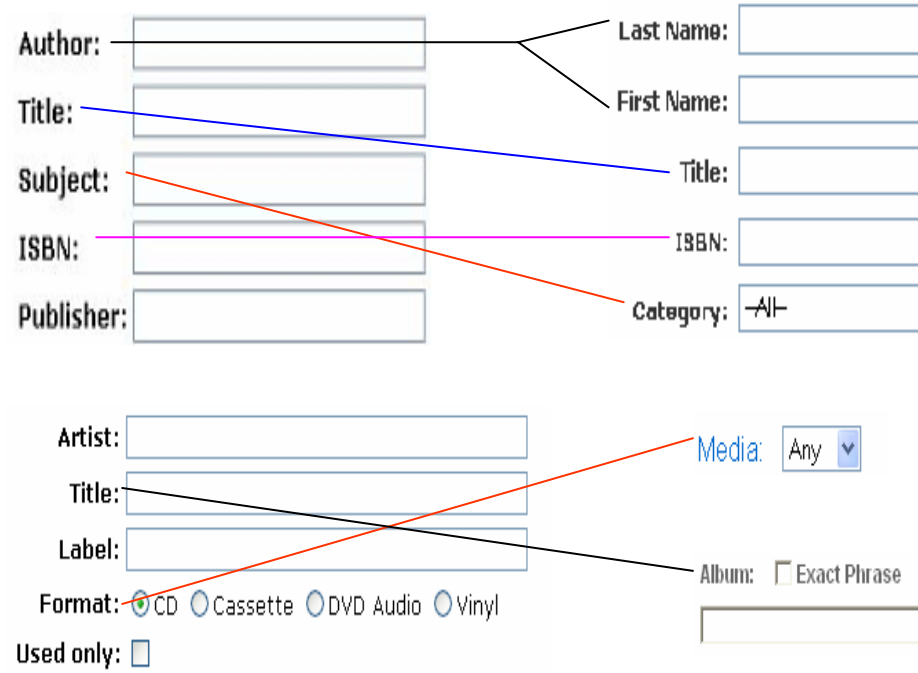


Source: Kevin Chen-Chuan Chang, CIDR 2005

→ can use statistical learning to derive mappings among different sources within same domain

Query-to-Source Mapping





Source: Kevin Chen-Chuan Chang, CIDR 2005

Deep Web search with MetaQuerier
http://metaquerier.cs.uiuc.edu/formext/

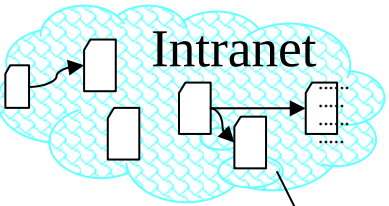
1.5 Intranet and Enterprise Search

Important differences to Web search:

- more professional query topics, more directed queries
- higher user skills, but also higher expectations
- fewer spam and low-quality pages
- fewer hyperlinks (endorsements)
- more meaningful metadata (e.g. author and date of document)
- more context knowledge about user and data
  (e.g. organizational units, location)
- opportunity for user (group) profiling

non-difference: intranet/enterprise data can also be huge !

System Architecture for Unstructured and Semistructured Information Management



integrate text & data,
and move up in information value chain

CMS

(semi-struct.) text doc.

struct. data-bases

information enrichment

Token-ization

NLP tagging (PoS etc.)

Named Entity Recognition

Record Linkage (Entity Reconciliation)

Relation extraction

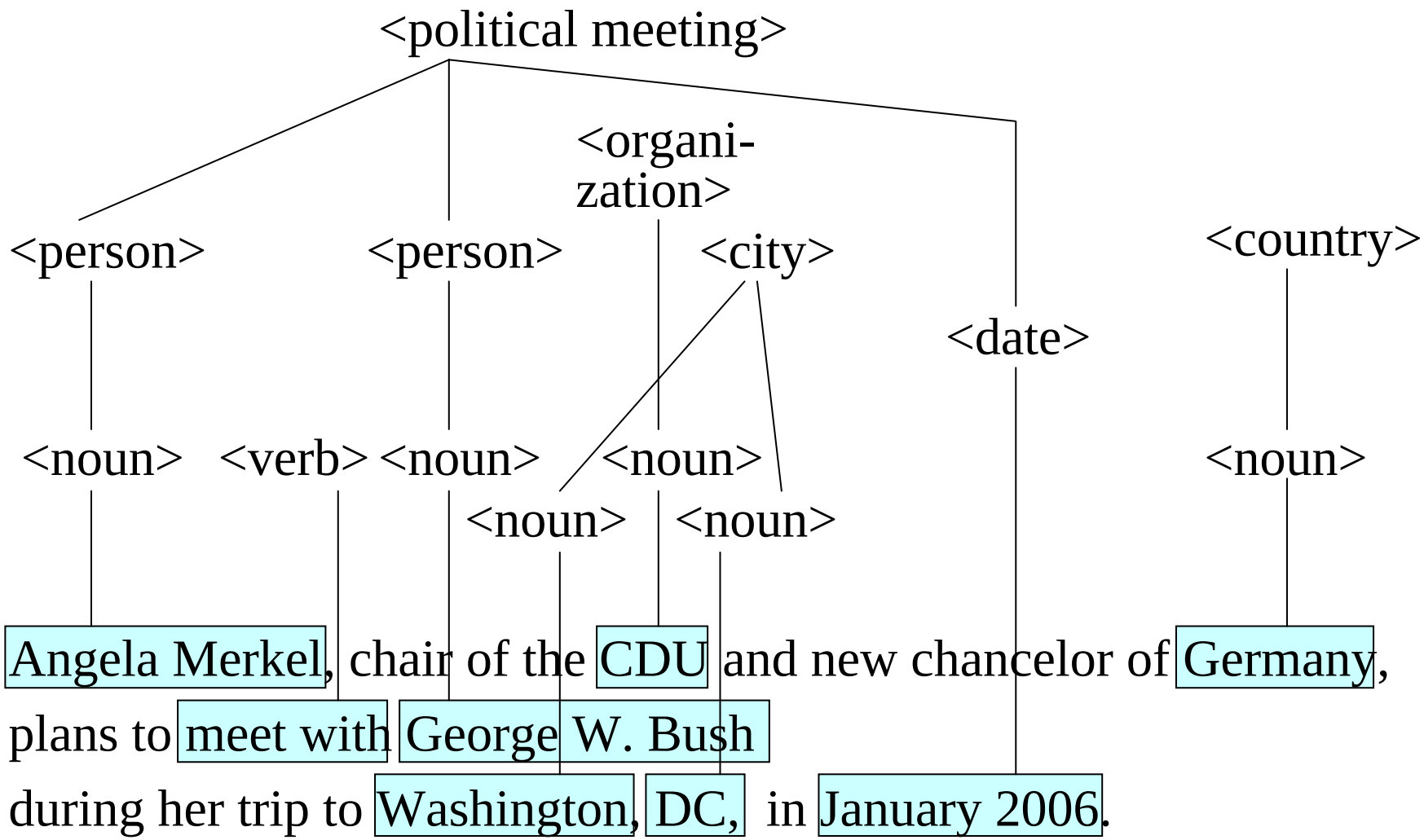
enriched (XML) data repository

dictionaries, ontologies

querying & mining for business analytics, customer support, health care, etc.

→ XML search

Information Enrichment Example



Information Extraction from Web Pages

ANNIE Output for http://en.wikipedia.org/wiki/Che_Guevarra

Annotation Key:
Person Location Organization Date Address Money Percent

>>/**/>/**/

Che Guevara

From Wikipedia, the free encyclopedia.

(Redirected from Che Guevarra)
Jump to: navigation , search

Che Guevara

Ernesto Rafael Guevara de la Serna (June 14 , 1928 [1] ? October 9 , 1967), commonly known as Che Guevara or el Che, was an Argentine -born Marxist revolutionary and Cuban guerrilla leader. Guevara was a member of Fidel Castro 's " 26th of July Movement " that seized power in Cuba in 1959 . After serving in various important posts in the new government, Guevara left Cuba in 1965 with the hope of fomenting revolutions in other countries, first in the Congo-Kinshasa (currently the Democratic Republic of the Congo) and later in Bolivia , where he was captured in a CIA -organized military operation. It is believed by some that the CIA wished to keep Guevara alive for interrogation but, after his capture in the Yuro ravine, he died at the hands of the Bolivian Army in La Higuera near Vallegrande on October 9 , 1967 . Testimony by various individuals who were participants in, or

Leading open-source tool: GATE/ANNIE
http://www.gate.ac.uk/annie/

1.6 Personalized Search and PIM

or

Personalization:

- query interpretation depends on personal interests and bias
- need to learn user-specific weights for multi-criteria ranking (relevance, authority, freshness, etc.)
- can exploit user behavior (feedback, bookmarks, query logs, click streams, etc.)

Personal Information Management (PIM):

- manage, annotate, organize, and search all your personal data
  - on desktop (mail, files, calendar, etc.)
  - at home (photos, videos, music, parties, invoices, tax filing, etc.) and in smart home with ambient intelligence

Query-Log and Click-Stream Sharing in Communities

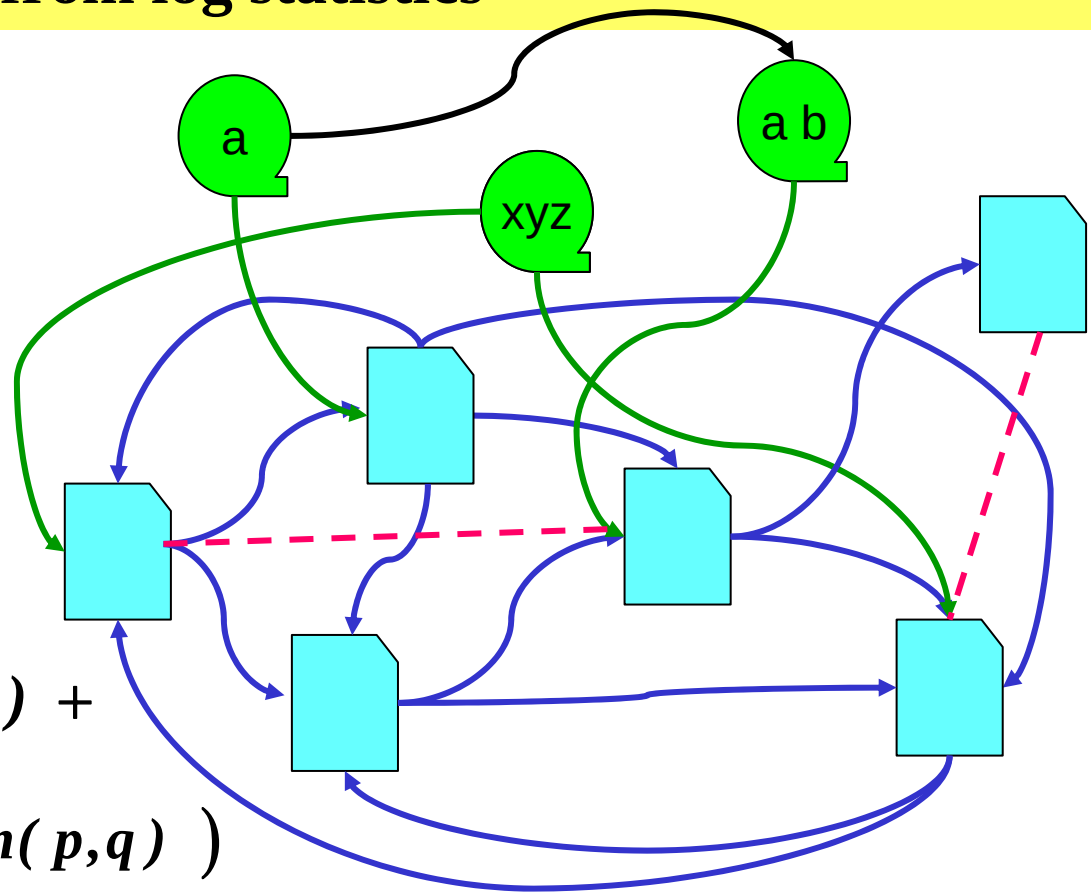
collect user queries, derive community profiles, and learn/adjust weights in multi-criteria ranking
http://ispy.ucd.ie/

Exploiting Query Logs and Click Streams

from PageRank: uniformly random choice of links + random jumps

to QRank: + query-doc transitions + query-query transitions
+ doc-doc transitions on implicit links (w/ thesaurus)
with probabilities estimated from log statistics

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



$$QR(q) = \varepsilon \cdot j(q) + (1-\varepsilon) \cdot \Big(\alpha \sum_{p \in explicitIN(q)} PR(p) \cdot t(p,q) + (1-\alpha) \sum_{p \in implicitIN(q)} PR(p) \cdot sim(p,q) \Big)$$

Preliminary Experiments

Setup:

70 000 Wikipedia docs, 18 volunteers posing Trivial-Pursuit queries
ca. 500 queries, ca. 300 refinements, ca. 1000 positive clicks
ca. 15 000 implicit links based on doc-doc similarity

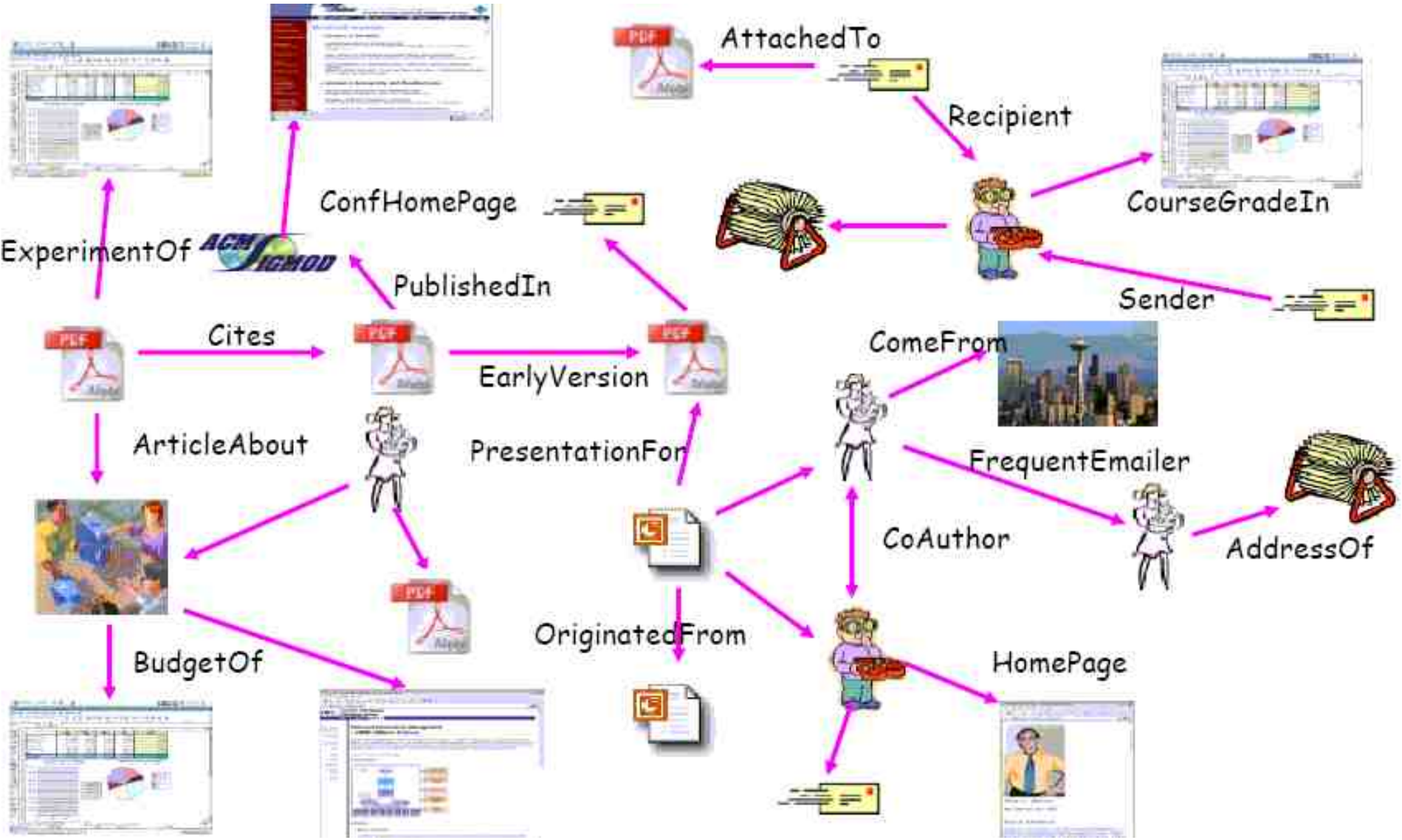
Results (assessment by blind-test users):

- **QRank top-10 result preferred over PageRank in 81% of all cases**
- **QRank has 50.3% precision@10, PageRank has 33.9%**

Untrained example query „philosophy“:

	PageRank	QRank
1.	Philosophy	Philosophy
2.	GNU free doc. license	GNU free doc. license
3.	Free software foundation	Early modern philosophy
4.	Richard Stallman	Mysticism
5.	Debian	Aristotle

Personal Information Management Challenge



Source: Alon Halevy, Semex Project, http://www.cs.washington.edu/homes/alon/

1.7 Peer-to-Peer (P2P) Search

P2P systems: decentralized, self-organizing, highly dynamic networks of loosely coupled, autonomous computers

Applications:

- Large-scale distributed computation (SETI, PrimeNumbers, etc.)
- **File sharing** (Napster, Gnutella, KaZaA, BitTorrent, etc.)
- IP telephony (Skype)
- Publish-Subscribe Information Sharing (Auctions, Blogs, etc.) with continuous queries (subscriptions) and alerting on updates
- Collaborative Work (Games, etc.)
- Collaborative Data Mining
- **(Collaborative) Web Search** (much harder than file search)

Why P2P Web Search

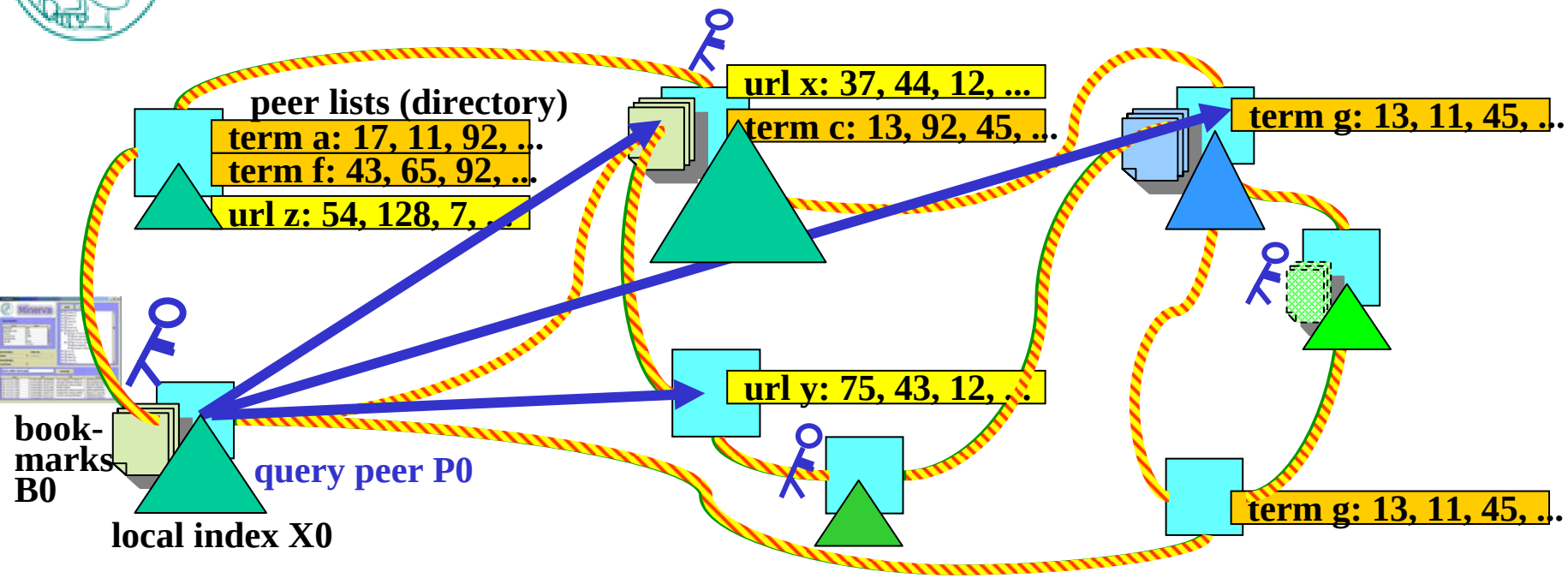
Objective: Self-organizing P2P Web Search Engines
with Google-or-better functionality

- Scalable & Self-Organizing Data Structures and Algorithms
  (DHTs, Semantic Overlay Networks, Epidemic Spreading, Distr. Link Analysis, etc.)

- Better Search Result Quality (Precision, Recall, etc.)
  - Powerful Search Methods for Each Peer
    (Concept-based Search, Query Expansion, Personalization, etc.)
  - Leverage Intellectual Input at Each Peer
    (Bookmarks, Feedback, Query Logs, Click Streams, Evolving Web, etc.)
  - Collaboration among Peers
    (Query Routing, Incentives, Fairness, Anonymity, etc.)

- Small-World Phenomenon
  Breaking Information Monopolies

Minerva System Architecture



Query routing aims to optimize benefit/cost driven by distributed statistics on peers' content similarity, content overlap, freshness, authority, trust, performability etc.

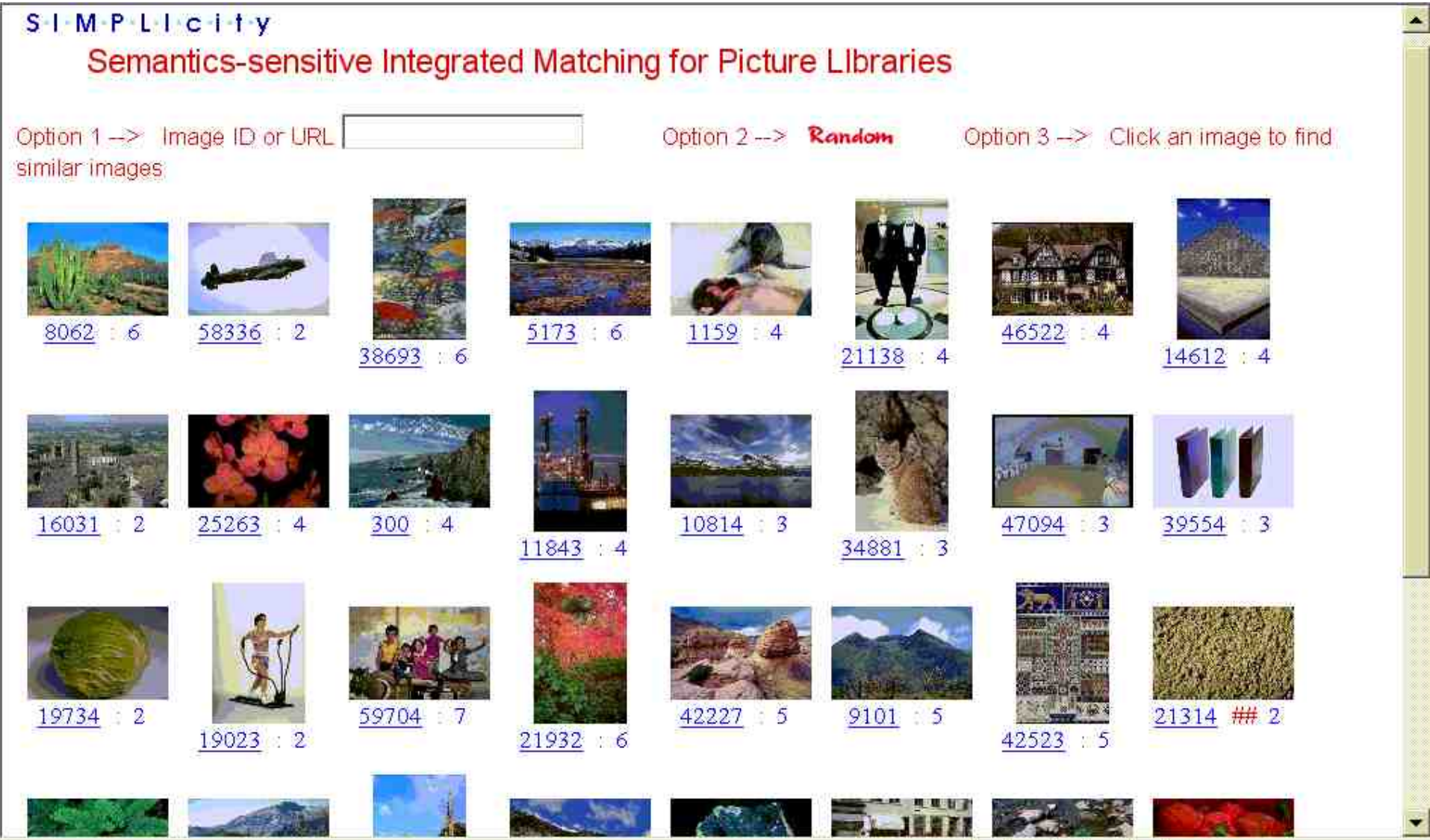
Dynamically precompute „good peers" to maintain a **Semantic Overlay Network** using random but biased graphs

1.8 Multimedia and NLP Search

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

Content-based Image Retrieval by Example (1)



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

Content-based Image Retrieval by Example (2)

Automatic Annotation of Images

S·I·M·P·L·I·c·i·t·y *a*–**LIP**

Automatic Linguistic Indexing of Pictures (wang.ist.psu.edu/IMAGE)

Random <- Click random to see more examples randomly selected from 60,000 images.
Or, click on a keyword below to search the computer-annotated image database by the selected keyword.

All keywords for the top 5 categories are shown. Words in **bold** are those top picks, selected by computer based on their statistical significances, for annotation. The computer is selecting keywords from **a dictionary of 600 automatically learned concepts**. Click here for more information about the project.

Computer Predictions-
grass Grand_canyon lion lizard animal mushroom wild_cat waterfall cloud lake rock mountain

Manual Category Annotation-
lizard animal rock

Computer Predictions-
horse grass barn_yard animal nature car rural plant flower man-made landscape

Manual Category Annotation-
Kenya Africa animal landscape people

Computer Predictions-
balloon sky ski city man-made snow mountain cloth building historical_building people landscape

Manual Category Annotation-
city historical_building building

Computer Predictions-
man-made sport car pill dining

Computer Predictions-
Toronto aviation rural scene grass plane England landscape city

Computer Predictions-
sun sky cloud dawn drink dusk orange food plant indoor

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

Natural-Language Question Answering

AnswerBus

who invented quantum theory Ask

Type in your question in English, French, Spanish, German, Italian or Portuguese:

Question:

who invented quantum theory

Possible answers: XML TXT

1. Shortly after quantum field theory was invented, people started trying to invent a quantum field theory of gravity.
2. Quantum field theory was invented to deal simultaneously with special relativity and quantum mechanics, the two greatest discoveries of early twentieth-century physics, but it has become increasingly important to many areas of physics.
3. Quantum field theory was invented to reconcile quantum mechanics with special relativity.
4. Planck himself in [7] explains how despite having invented quantum theory he did not understand it himself at first:- I tried immediately to weld the elementary quantum of action somehow in the framework of classical theory.
5. Meanwhile, there is no doubt that quantum mechanics is the most successful theory of physical phenomena yet invented by the human mind.
6. It has not been reached?not by quantum theory, not by special or general relativity, not by anything invented since.
7. Austrian physicist Erwin Schr dinger, who, like Albert Einstein, never really believed in quantum theory, invented the story of a cat, now named after him, to illustrate how absurd the situation is.

find compact text passages for question answering
http://answerbus.coli.uni-sb.de/index.shtml

Additional Literature (1)

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

performance evaluation 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://inex.is.informatik.uni-duisburg.de/
- KDD Cup, http://www.kdnuggets.com/datasets/kddcup.html and http://kdd05.lac.uic.edu/kddcup.html
- Language-Independent Named-Entity Recognition, www.cnts.ua.ac.be/conll2003/ner/

Additional Literature (2)

Crawling, storage, and server management:

- S. Brin, L. Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine, WWW 1998
- S. Ghemawat, H. Gobioff, S.-T. Leung: The Google File System, SOSP 2003
- P. Boldi, B. Codenotti, M. Santini, S. Vigna. Ubicrawler: A Scalable Fully Distributed Web Crawler, Software: Practice & Experience, 34(8):711-726, 2004.
- A. Heydon, M. Najork: Mercator: A Scalable, Extensible Web Crawler, WWW 1999.
- V. Shkapenyuk, T. Suel: Design and Implementation of a High-Performance Distributed Web Crawler, ICDE 2002
- X. Long, T. Suel: Three-level caching for efficient query processing in large Web search engines, WWW 2005

Web structure, size, dynamics:

- D.E. Rose, D. Levinson: Understanding User Goals in Web Search, WWW 2004
- A. Gulli, A. Signorini: The Indexable Web is More Than 11.5 Billion Pages, WWW 2005
- A. Ntoulas, J. Cho, C. Olston: What‘s New on the Web? The Evolution of the Web from a Search Engine Perspective, WWW 2004
- D. Fetterly, M. Manasse, M. Najork: Spam, Damn Spam, and Statistics, WebDB 2004
- D. Fetterly, M. Manasse, M. Najork, J. Wiener: A Large-Scale Study of the Evolution of Web Pages, WWW 2003
- A.Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins,
- J.L. Wiener: Graph structure in the Web, WWW 2000
- D. Donato et al.: Mining the Inner Structure of the Web Graph, WebDB 2005 Web graph data and tools, http://webgraph.dsi.unimi.it/
- Center for Complex Network Research, http://www.nd.edu/~networks/
- Search Engine Watch, http://searchenginewatch.com/

Additional Literature (3)

Geo- and time-aware search:

- A. Markowetz, Y.-Y. Chen, T. Suel, X. Long, B. Seeger: Design and Implementation of a Geographic Search Engine, WebDB 2005
- D. Ancona. J.Frew, G.Janée, and D.Valentine: Accessing the Alexandria Digital Library from Geographic Information Systems, JCDL 2004
- J. Ding, L. Gravano, N. Shivakumar: Computing Geographical Scopes of Web Resources, VLDB 2000
- K. Berberich, M. Vazirgiannis, G. Weikum: Time-aware Authority Ranking, to appear in Internet Mathematics Journal

Deep Web search:

- Kevin Chen-Chuan Chang, Bin He, Zhen Zhang: Toward Large Scale Integration: Building a MetaQuerier over Databases on the Web, CIDR 2005
- B. He, M. Patel, Z. Zhang, K. C.-C. Chang : Accessing the Deep Web: a Survey, Communications of the ACM, 2006
- Luciano Barbosa, Juliana Freire: Searching for Hidden-Web Databases, WebDB 2005

Intranet and enterprise search:

- IBM Systems Journal 43(3), 2004, Special Issue on Unstructured Information Management, http://www.research.ibm.com/journal/sj43-3.html
- R. Fagin, R. Kumar, K.S. McCurley, J. Novak, D. Sivakumar, J.A. Tomlin, D.P. Williamson: Searching the Workplace Web, WWW 2003
- Bjorn Olstad: Why Search Engines are used increasingly to Offload Queries from Databases, Keynote, VLDB 2005, http://www.vldb2005.org/program/slides/tue/s1-olstad.ppt
- Aleksander Ohrn: Contextual Insight in Search: Enabling Technologies and Applications, Tutorial, VLDB 2005, http://www.vldb2005.org/program/slides/wed/s1366-ohrn.ppt

Additional Literature (4)

Personalized search and PIM:

- J. Luxenburger, G. Weikum: Query-Log Based Authority Analysis for Web Information Search, WISE 2004
- E. Balfe, B. Smyth: An Analysis of Query Similarity in Collaborative Web Search, ECIR 2005
- E. Balfe, B. Smyth: Improving Web Search through Collaborative Query Recommendation, ECAI 2004: 268-272
- T. Mitchell: Computer Workstations as Intelligent Agents, Keynote, SIGMOD 2005, http://www.cs.cmu.edu/~tom/
- G. Bell: MyLifeBits: a Memex-Inspired Personal Store; Another TP Databa, Keynote, SIGMOD 2005, http://research.microsoft.com/users/GBell/
- X. Dong, A. Halevy: A Platform for Personal Information Management and Integration, CIDR 05
- X. Dong, A. Halevy, Jayant Madhavan: Reference Reconciliation in Complex Information Spaces, SIGMOD 2005

P2P search and collaboration:

- M. Bender, S. Michel, P. Triantafillou, G. Weikum, C. Zimmer: Improving Collection Selection with Overlap Awareness in P2P Search Engines, SIGIR 2005
- J.X. Parreira, G. Weikum: JXP: Global Authority Scores in a P2P Network, WebDB 2005
- J. Zhang, T. Suel: Efficient Query Evaluation on Large Textual Collections in a Peer-to-Peer Environment, Int.Conf. on Peer-to-Peer Computing, 2005
- F. M. Cuenca-Acuna, C. Peery, R. P. Martin, T. D. Nguyen: PlanetP: Using Gossiping to Build Content Addressable Peer-to-Peer Information Sharing Communities, HPDC 2003
- Christos Tryfonopoulos, Stratos Idreos, Manolis Koubarakis: Publish/Subscribe Functionality in IR Environments using Structured Overlay Networks, SIGIR 2005

Additional Literature (5)

Multimedia and NLP search:

- J.Z. Wang, J. Li, G. Wiederhold: SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries, IEEE Trans. on Pattern Analysis and Machine Intelligence 23(9), 2001
- J. Li, J.Z. Wang: Automatic linguistic indexing of pictures by a statistical modeling approach, IEEE Transactions on Pattern Analysis and Machine Intelligence 25(9), 2003
- M. Ortega-Binderberger, S. Mehrotra: Relevance feedback techniques in the MARS image retrieval system. Multimedia Syst. 9(6), 2004
- J. Fauqueur, N. Boujemaa: Region-based image retrieval: fast coarse segmentation and fine color description. J. Vis. Lang. Comput. 15(1), 2004
- A. Natsev, R. Rastogi, K. Shim: WALRUS: A Similarity Retrieval Algorithm for Image Databases, IEEE Trans. Knowl. Data Eng. 16(3), 2004
- Y. Zhu, D. Shasha: Warping Indexes with Envelope Transforms for Query by Humming, SIGMOD 2003
- E. Agichtein, S. Lawrence, L. Gravano: Learning to find answers to questions on the Web. ACM TOIT 4(2): 129-162 (2004)
- Ganesh Ramakrishna, Soumen Chakrabarti, Deepa Paranjpe, Pushpak Bhattacharya: Is question answering an acquired skill? WWW 2004
- Zhiping Zheng: Question Answering Using Web News as Knowledge Base, EACL 2003