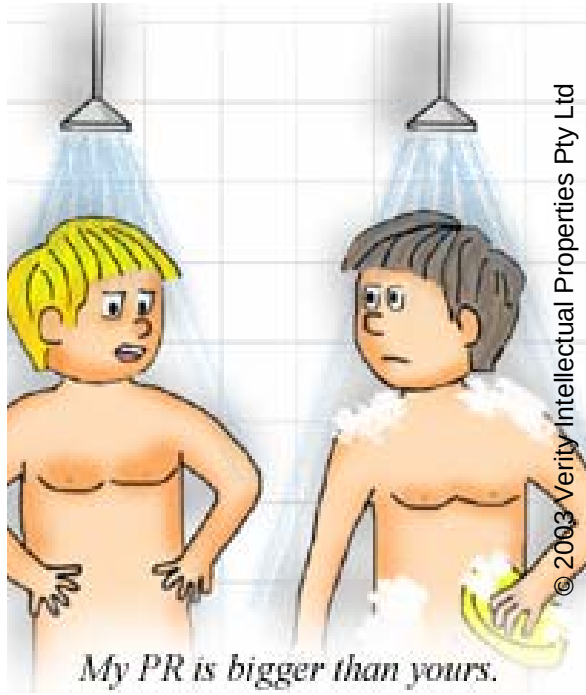
max planck institut
informatik

Next-Generation (Peer-to-Peer) Search Engines

Gerhard Weikum (weikum@mpi-inf.mpg.de)

Google is Great



- 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

- *professors from Saarbruecken who teach DB or IR and have projects on XML*
- *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*

apps in customer support, business analytics, health care, etc.
+ multilingual/multicultural, personalized/contextual, multimedia, etc.

What Are We Missing?

for Advanced Information Requests by „Power Users“
(librarians, market analysts, scientists, students, etc.)

- ***background knowledge***
  → ontologies & thesauri, statistics, continuous learning

- ***(semi-)structured and „semantic“ data***
  → XML, info extraction, (cont.) annotation & classification

- ***humans in the loop***
  → collaboration, recommendation, peers

- ***context awareness***
  → personalization, geo & time, user behavior, rea

→ Peer-to-Peer (P2P) Search Engines:
Wisdom of the Crowds !

Why Peer-to-Peer Search Engines?

Vision: 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.)

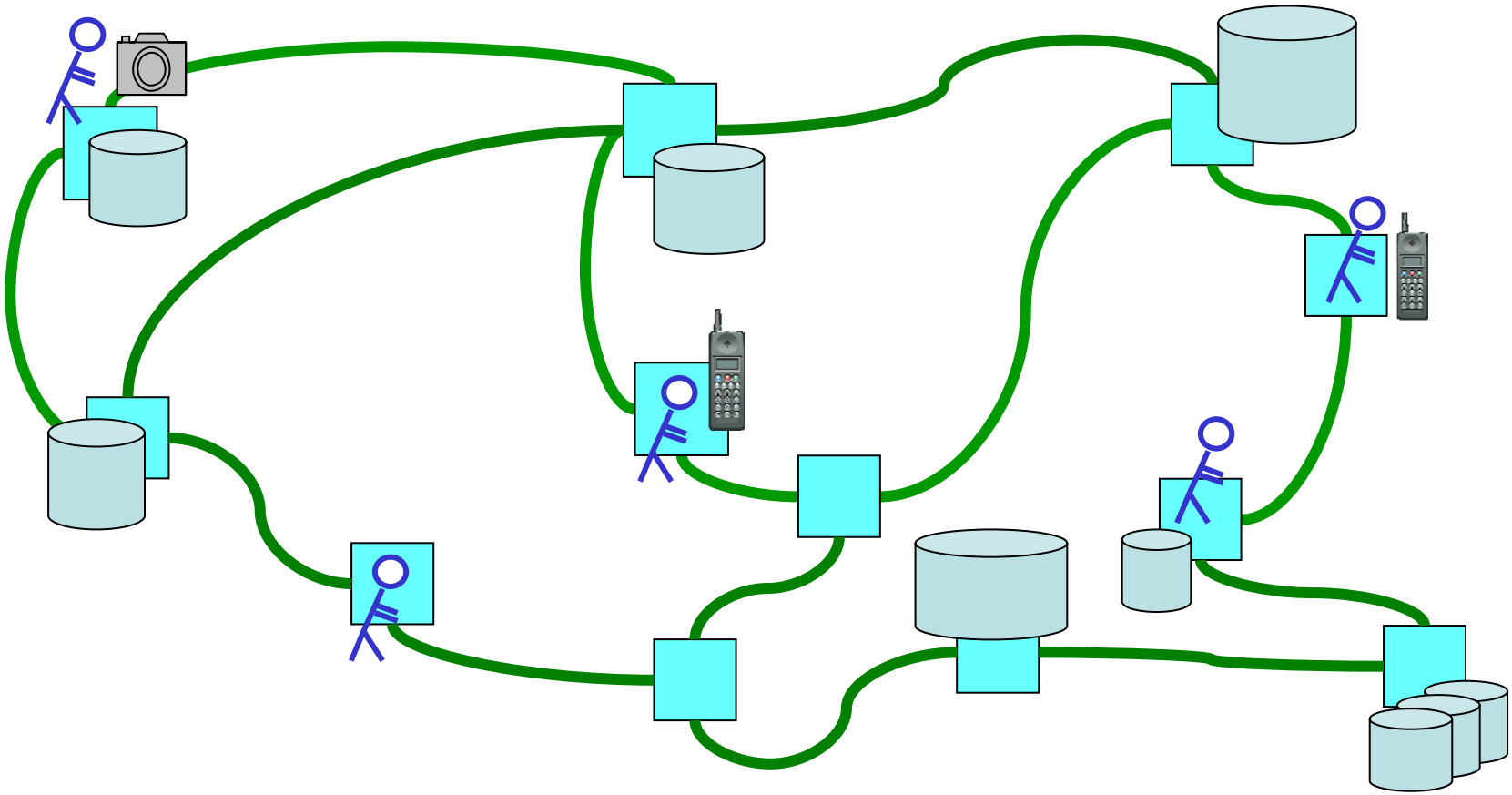
- Better Search Result Quality (Precision, Recall, etc.)
  - Powerful Search Methods for Each Peer
    (Concept-based Search, Query Expansion, Personalization, etc.)
  - Leverage Intellectual Input at Each Peer
    (Bookmarks, Feedback, Query Logs, Click Streams, Evolving Web, etc.)
  - Collaboration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

- Small-World Phenomenon
  Breaking Information Monopolies

foundations pursued in IP DELIS, application to DLs explored in NoE DELOS

P2P Architecture for DLs and DL Users

Self-organizing overlay networks for info sharing, PubSub, recommendations, search, routing (e.g. BitTorrent, Skype, etc.)



Peers: • **DLs, Citation Servers, Annotation Servers, Image Repositories, Public Databases, Web Archives, News Feeds, Blogs, etc.**

• **Users, Mobile Devices, etc.**

Outline

- ✓ Motivation and Strategic Direction
- Example: Richer Data
- Example: Personalization
- Conclusion

Query Expansion in 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:

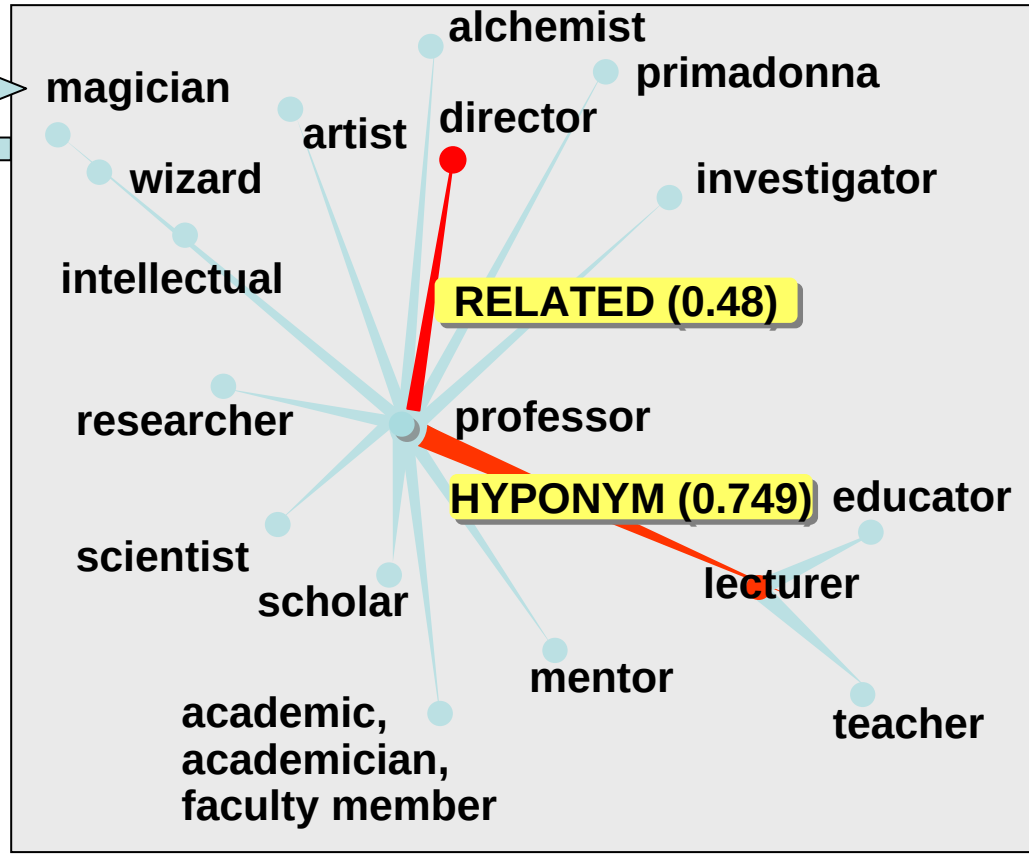
(professor lecturer (0.749) scholar (0.71) ...)
and ((course class (1.0) seminar (0.84) ...)
= („IR“ „Web search“ (0.653) ...))

Efficient top-k search with dynamic expansion

better recall, better mean precision for hard queries

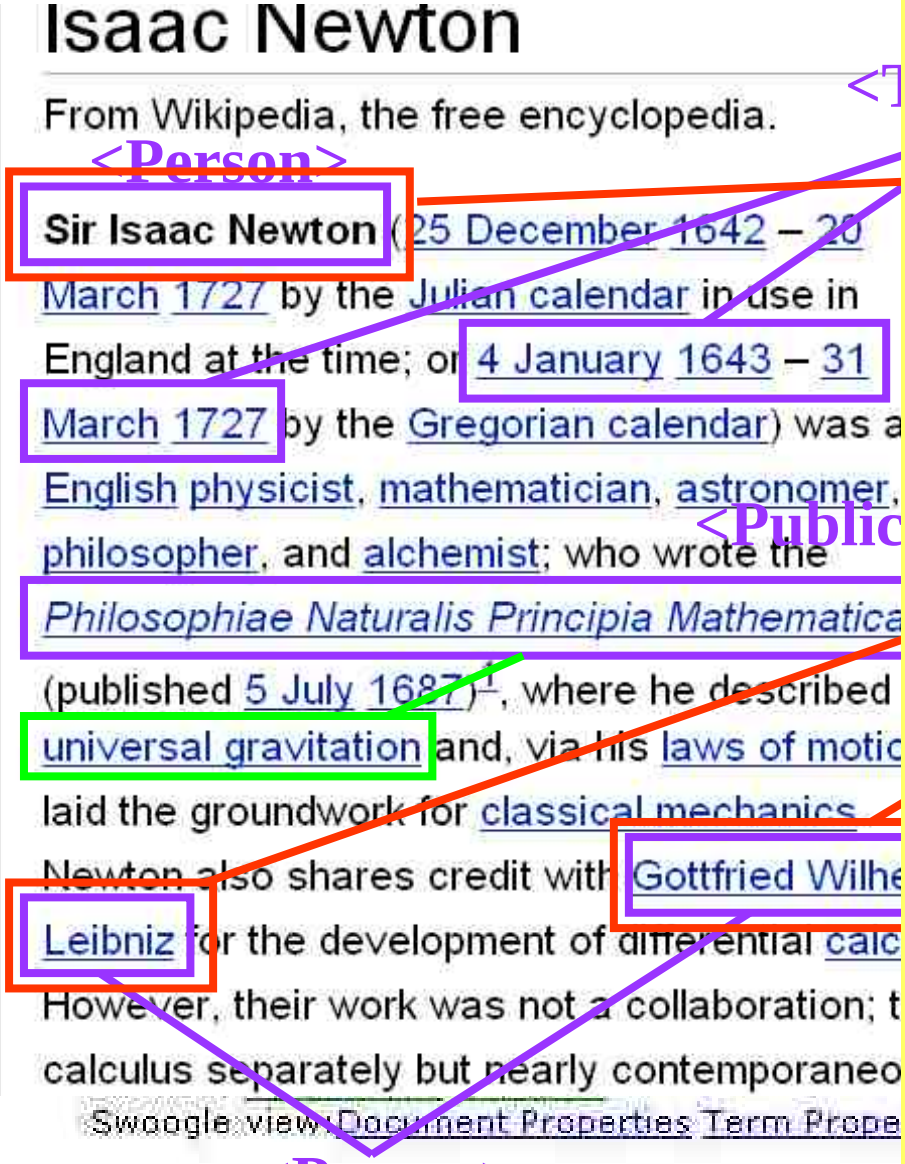
Thesaurus/Ontology:

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 retrieval !**

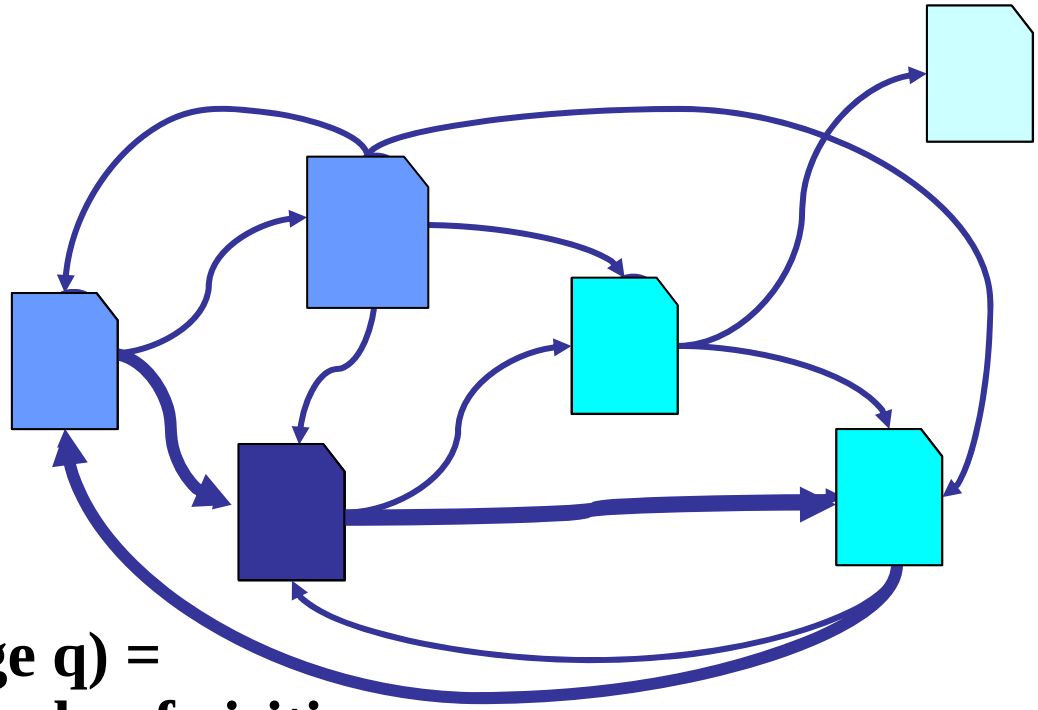
Outline

- ✓ Motivation and Strategic Direction
- ✓ Example: Richer Data
- Example: Personalization
- Conclusion

Google's PageRank Reviewed

from PageRank: uniformly random choice of links + random jumps

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



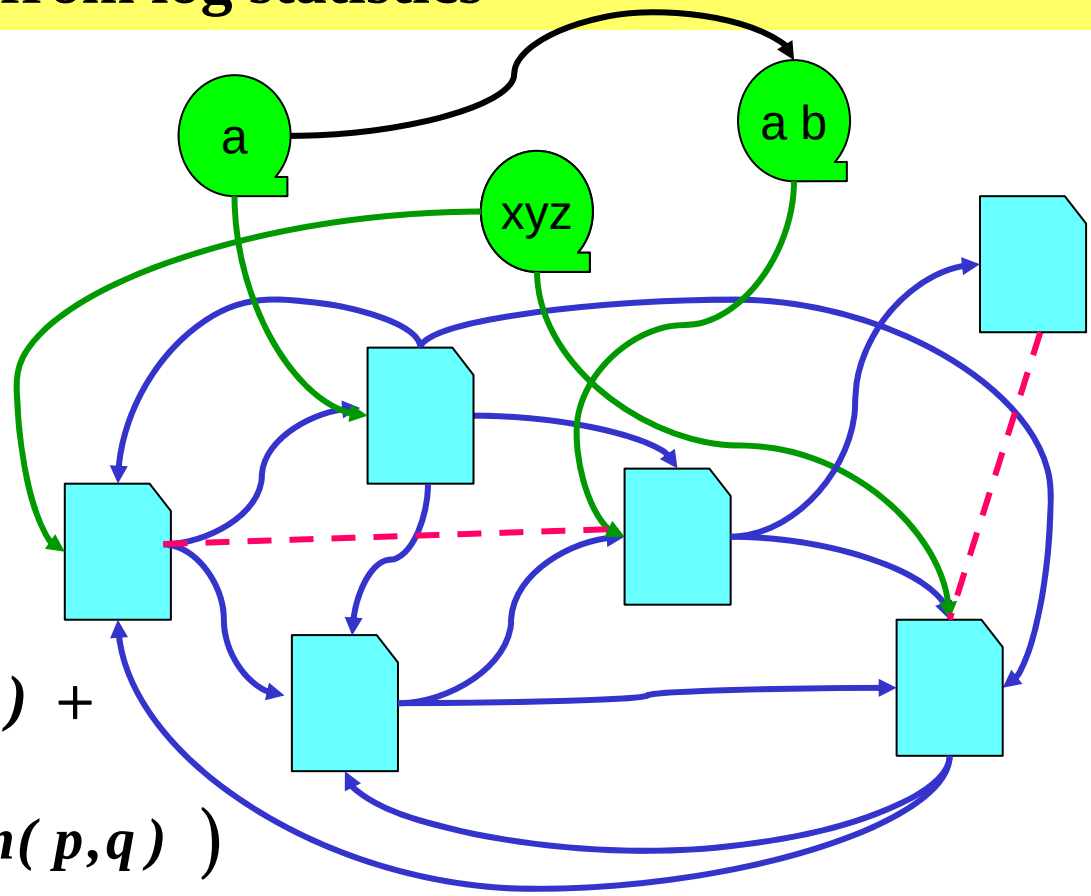
Authority (page q) = stationary prob. of visiting q

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

Outline

- ✓ Motivation and Strategic Direction
- ✓ Example: Richer Data
- ✓ Example: Personalization
- Conclusion

Conclusion

P2P search engines have great potential:

- harness local resources for power search engine
- rich models for content extraction, annotation, summarization, and indexing of text, images, speech, audio&video, feeds, portals
- customization and personalization
- collaboration & recommendation networks with other peers
- naturally fits with mobile clients and context awareness
- naturally gears for rich cognitive model of user behavior
- no monopoly, no central profiling or bias
- great benefit for European society, economy, science
- business applications in intranets, communities, web archives, search embedded in business intelligence, mobile apps, etc.