max planck institut
informatik

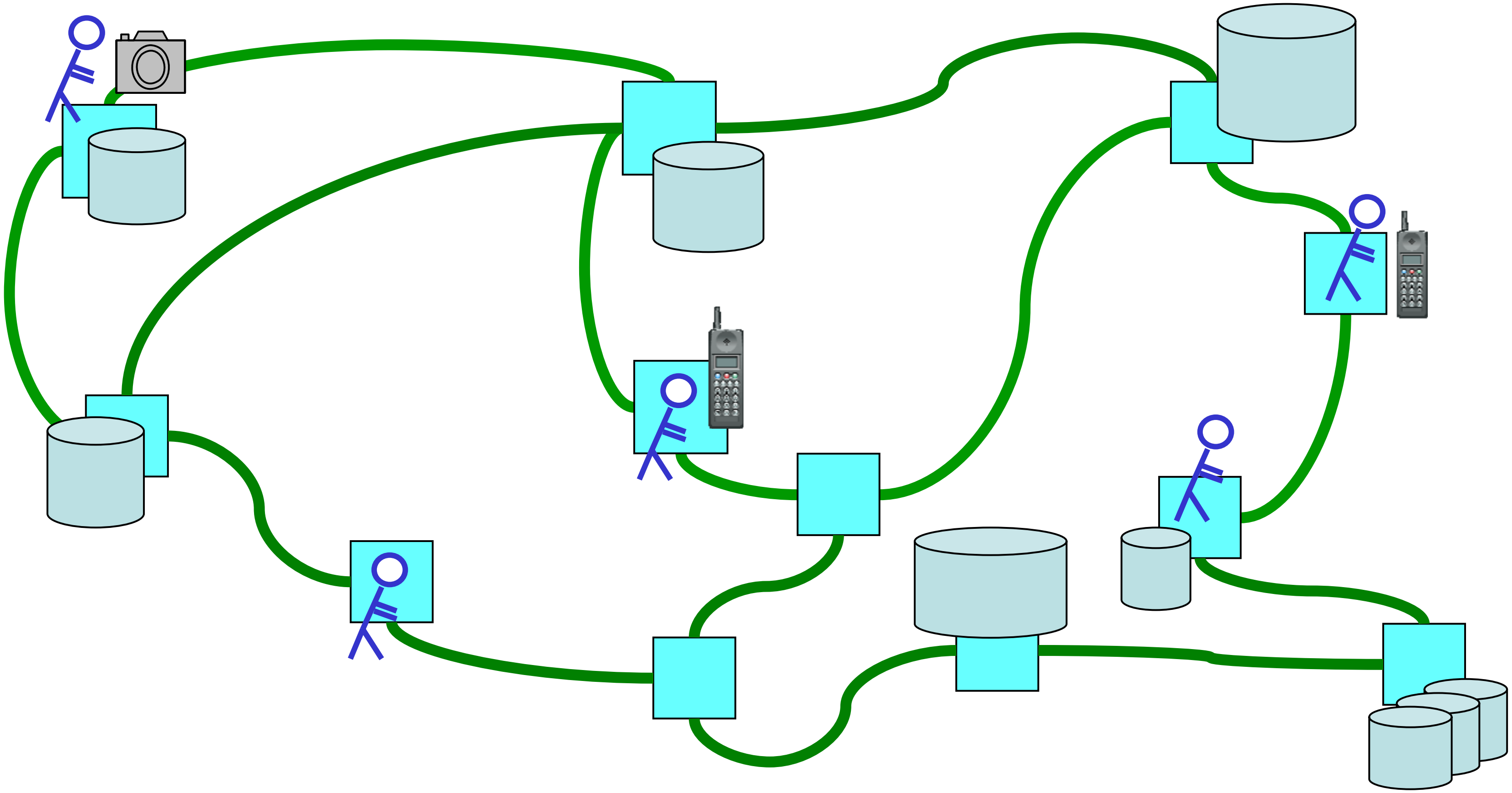
DELOS Task 2.8:
Personalized Query Routing in Peer-to-Peer Federations of Digital Libraries

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

DELOS
Network of
Excellence on
Digital
Libraries

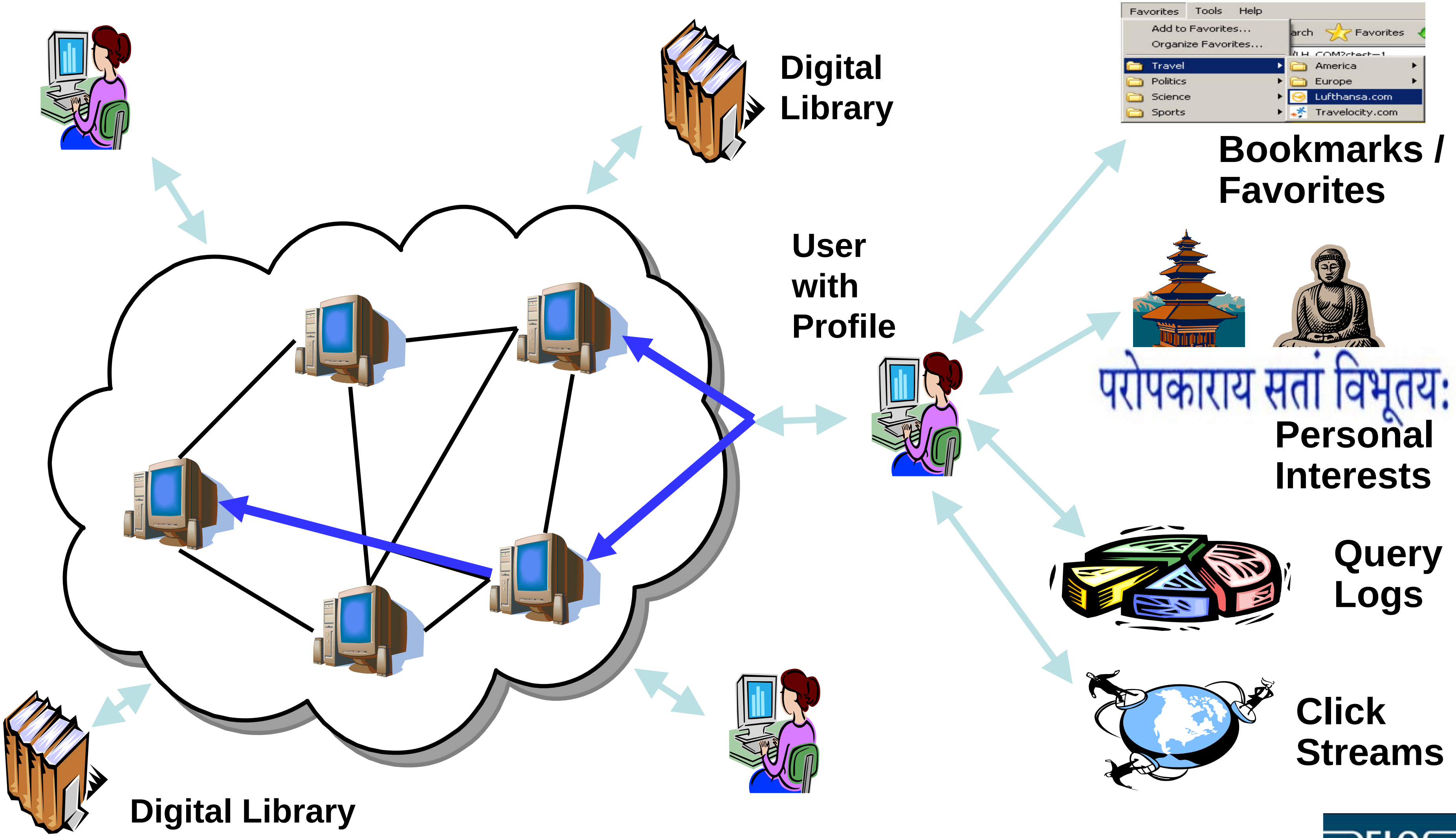
P2P Architecture for DLs and DL Users

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.**

Opportunities and Challenges of Personalized P2P Search



Digital Library

Bookmarks / Favorites

User with Profile

Personal Interests

Query Logs

Click Streams

Digital Library

Task 2.8: Goal and Partners

Goal:

models and strategies for personalized query routing (selecting peers based on user profile & history)

Partners and their Expertise:

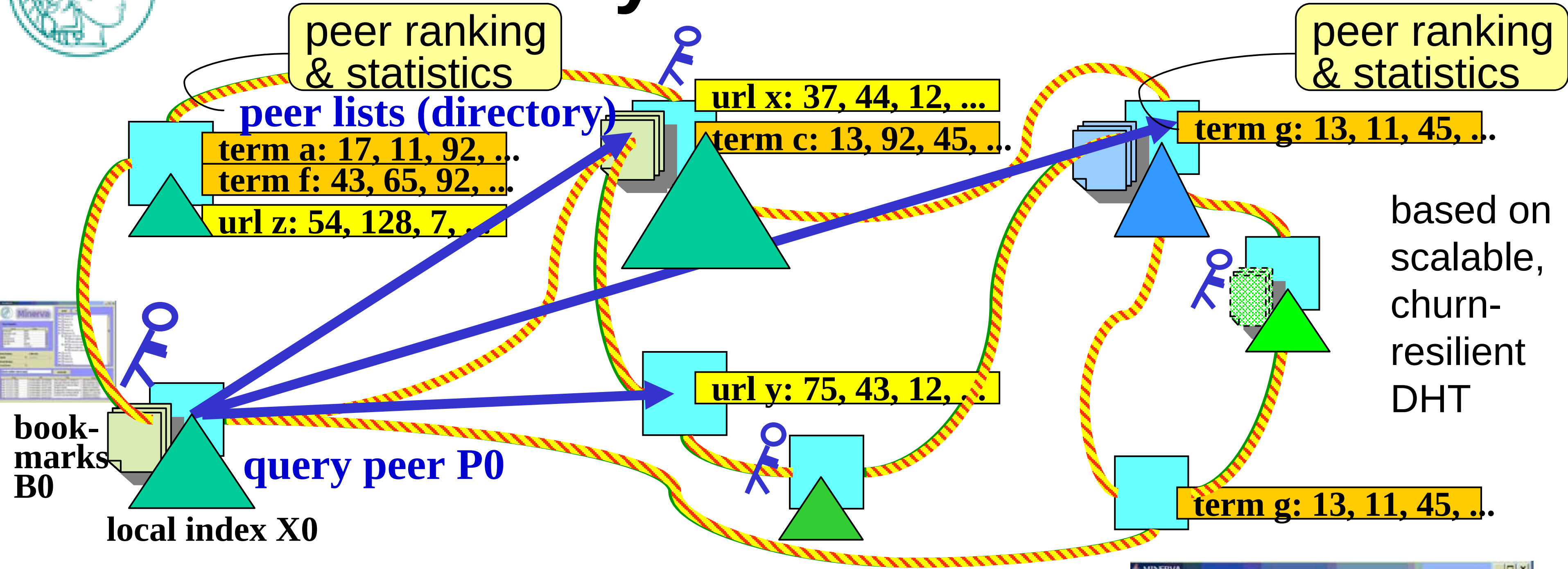
- **Max-Planck Institute for Informatics Saarbrücken (Gerhard Weikum):**
  P2P Web search
- **National University of Athens (Yannis Ioannidis):**
  user profiles, preference queries
- **University for Health Sciences Innsbruck (Hans-Jörg Schek):**
  relevance feedback, e-health apps
- **University of Duisburg-Essen (Norbert Fuhr):**
  P2P IR, DL agents
- **Masaryk University Brno (Pavel Zezula):**
  distributed similarity search
- **ETH Zurich (Donald Kossmann):**
  scalable, personalized PubSub, desktop search

Outline

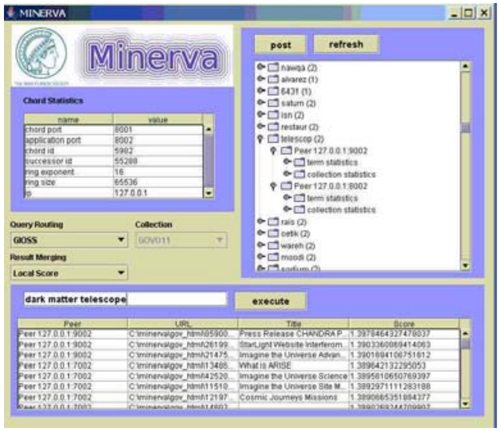
✓ Motivation and Research Direction

- P2P Search Engine
- Query Routing
- Conclusion

Minerva System Architecture



based on scalable, churn-resilient DHT



Chord Statistics

name	value
chord port	8001
application port	8002
chord id	5982
successor id	55288
ring exponent	16
ring size	65536
ip	127.0.0.1

Query Routing: GlOSS

Collection: GOV011

Result Merging: Local Score

- nawqa (2)
- alvarez (1)
- 6431 (1)
- saturn (2)
- isn (2)
- restaur (2)
- telescop (2)
  - Peer 127.0.0.1:9002
    - term statistics
    - collection statistics
  - Peer 127.0.0.1:8002
    - term statistics
    - collection statistics
- rais (2)
- cetik (2)
- wareh (2)
- moodi (2)
- sodium (2)

dark matter telescope

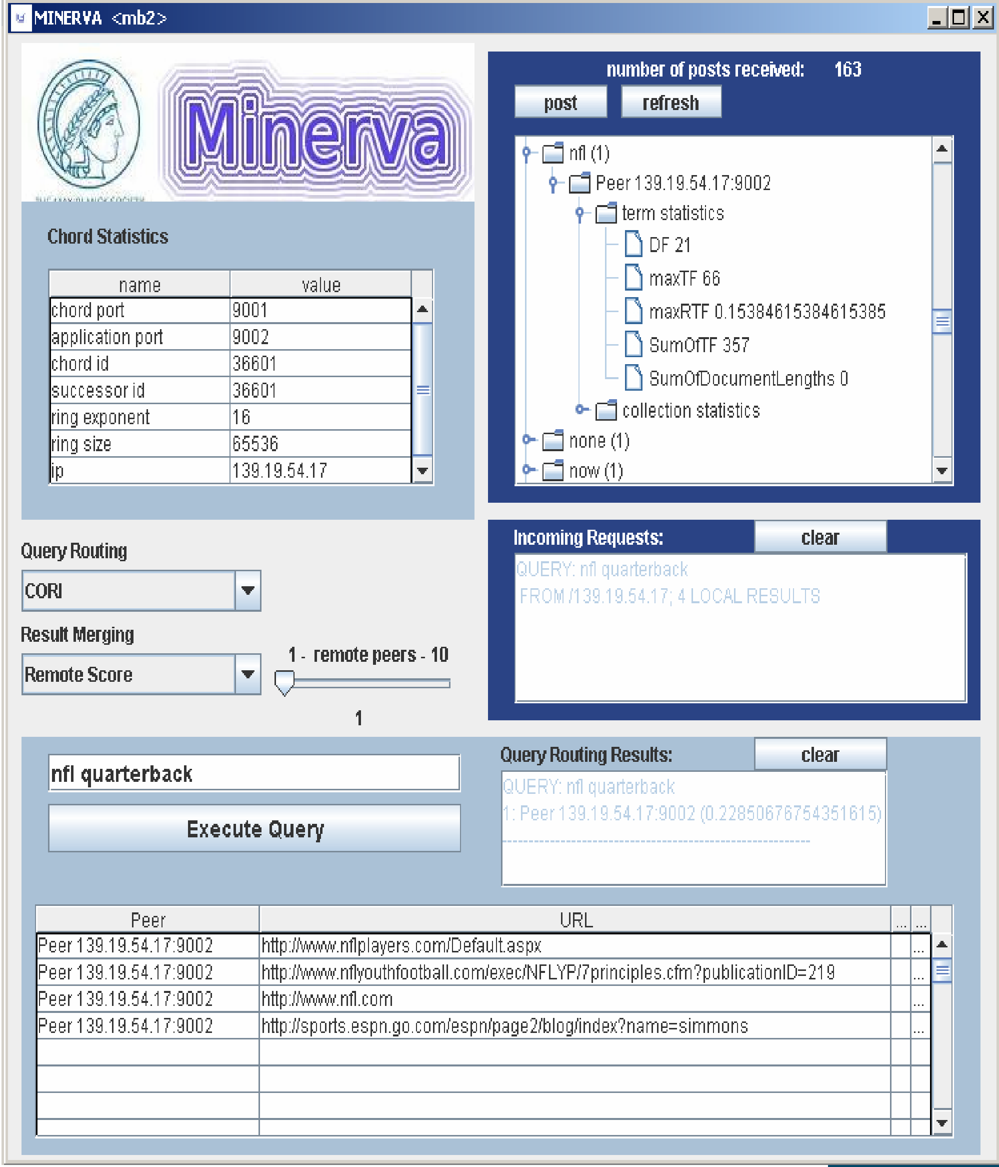
Peer	URL	Title	Score
Peer 127.0.0.1:9002	C:\minerva\gov_html\85900...	Press Release CHANDRA P...	1.3978464327478037
Peer 127.0.0.1:9002	C:\minerva\gov_html\26199...	StarLight Website Interferom...	1.3903360069414063
Peer 127.0.0.1:9002	C:\minerva\gov_html\21475...	Imagine the Universe Advan...	1.3901694106751612
Peer 127.0.0.1:7002	C:\minerva\gov_html\13485...	What is ARISE	1.3896421322950530
Peer 127.0.0.1:7002	C:\minerva\gov_html\42520...	Imagine the Universe Science	1.3895810650769397
Peer 127.0.0.1:7002	C:\minerva\gov_html\11510...	Imagine the Universe Site M...	1.3892971111283188
Peer 127.0.0.1:7002	C:\minerva\gov_html\12197...	Cosmic Journeys Missions	1.3890665351884377

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

Minerva at Work

- ✓**Peers Registering with MINERVA**
  - ✓Join DHT-style directory and inspect system status
  - ✓Post statistical metadata about local index
  - ✓Inspect metadata of other peers
- ✓**Query Routing and Processing with MINERVA**
  - ✓Enter keyword query
  - ✓Gather metadata from distributed directory to perform Query Routing
  - ✓Execute query at selected peers using top-k query execution strategies
- ✓**Query Result Merging and Display**
  - ✓Merge results into single result list at querying peer
  - ✓Click on query results to view (cached copies of) web pages



number of posts received: 163

nfl (1)
Peer 139.19.54.17:9002
term statistics
DF 21
maxTF 66
maxRTF 0.15384615384615385
SumOfTF 357
SumOfDocumentLengths 0
collection statistics
none (1)
now (1)

Chord Statistics

name	value
chord port	9001
application port	9002
chord id	36601
successor id	36601
ring exponent	16
ring size	65536
ip	139.19.54.17

Query Routing: CORI

Result Merging: Remote Score

1 - remote peers - 10

1

Incoming Requests:

QUERY: nfl quarterback
FROM /139.19.54.17; 4 LOCAL RESULTS

nfl quarterback

Query Routing Results:

QUERY: nfl quarterback
1: Peer 139.19.54.17:9002 (0.22850676754351615)

Peer	URL	...	...
Peer 139.19.54.17:9002	http://www.nflplayers.com/Default.aspx		...
Peer 139.19.54.17:9002	http://www.nflyouthfootball.com/exec/NFLYP/7principles.cfm?publicationID=219		...
Peer 139.19.54.17:9002	http://www.nfl.com		...
Peer 139.19.54.17:9002	http://sports.espn.go.com/espn/page2/blog/index?name=simmons		...

max planck institut
informatik

DELOS
Network of Excellence on Digital Libraries

Outline

✓ Motivation and Research Direction

✓ P2P Search Engine

- Query Routing

- Conclusion

Quality&Overlap-Aware Query Routing [SIGIR'05]

Select peers with highest benefit/cost ratio where

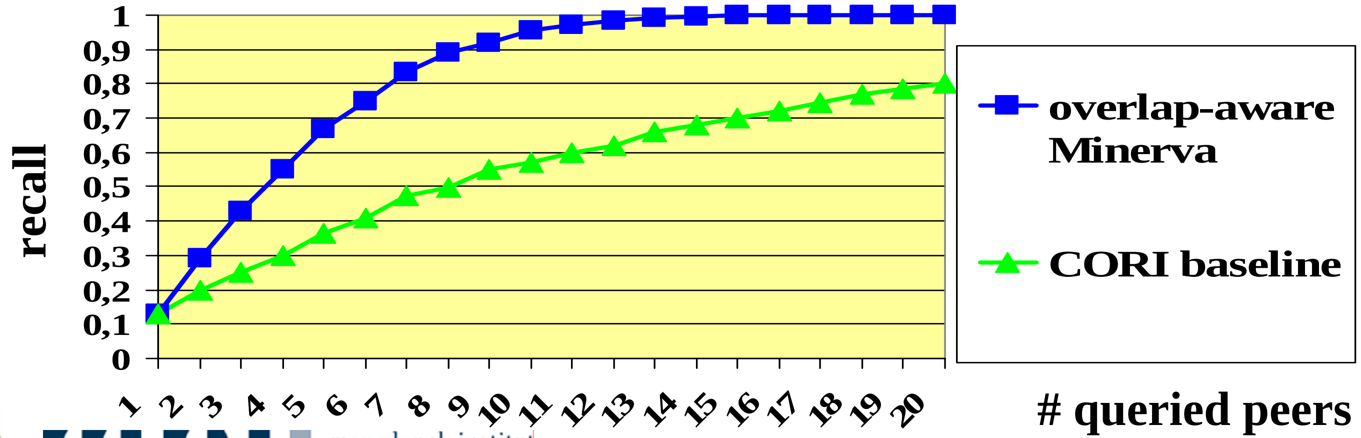
- benefit(Pi) ~ sim (X0, Xi) and ~ 1/overlap(X0, Xi)
  or using bookmarks B0, Bi for personalization & efficiency
- cost(Pi) ~ estimated response time or communication costs

precompute sim: $KL(X0, Xi) := \sum_{terms\ x} freq(x, X0)\ \log \frac{freq(x, X0)}{freq(x, Xi)}$

estimate overlap by Bloom filters, hash sketches, or MIPs

Experiments:

based on 100 .Gov partitions (1.25 Mio. docs), assigned to 50 peers,
with each peer holding 10 partitions and 80% overlap for P_i, P_{i+1}
with 50 TREC-2003 Web queries, e.g.: „juvenile delinquency“



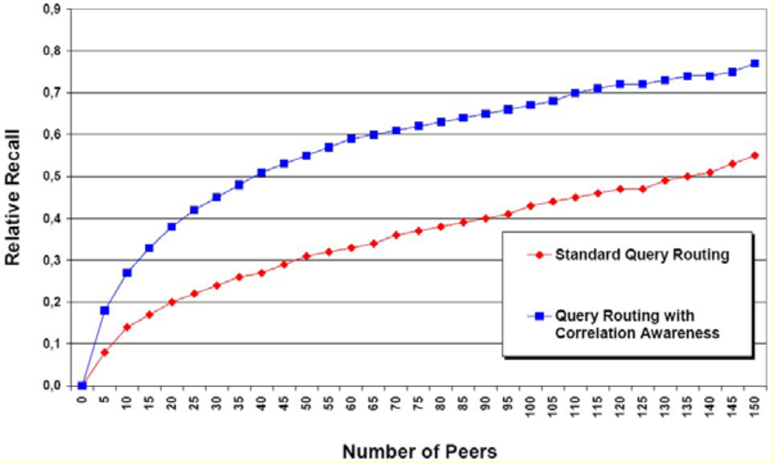
Considering Term Correlations [IPTPS'06]

Problem: DHT-based Per-Term Directory loses term correlations such as „Michael Jordan“ or „Native American Music“

Solution:

- peers perform frequent-itemset mining on local query log
- correlated termsets posted to all single-term directory peers
- directory peers collect postings for termsets from all peers
- query routed to single-term peers, evaluated over max. termsets
- all communication piggybacked on normal traffic, no extra cost

experiments based on 750 peers
with .Gov partitions,
running expanded queries
from TREC-2003 Web track;
examples:
„marijuana legalization drug abuse ...“,
„wireless communication broadcasting“



Distributed Similarity Search in Metric Spaces

Problem:

Scalable distributed indexing of data objects for kNN queries with metric distances satisfying triangle inequality dist(x,z) ≤ dist(x,y) + dist(y,z)

Approach: [Delos 2005]

- embed data objects into distance-preserving vector space
- map kNN queries into range queries
- index by dynamic partitioning across peers of DHT

Example: Edit Distance

query q: *Mex Plank Institute*
should be corrected into
query q‘: *Max Planck Institut*
based on P2P directory

and then submitted
to P2P search
(joint work MPII & MUNI)

Continuous Queries in P2P Publish-Subscribe

IR (Information Retrieval):
best results for
one-time query

vs.

IF (Information Filtering):
alerting about new docs
that match standing query

State-of-the-art IF considers only exact matches
and has only coarse-grained topics for personalization

Challenge (work in progress):
Approximate IF
should alert the user about vague matches
and may miss some docs with low probability
for better P2P scalability and churn-resilience,
and can support fine-grained personalization

Outline

✓ Motivation and Research Direction

✓ P2P Search Engine

✓ Query Routing

- 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

Query routing is the key issue in P2P search

Task 2.8: 6 partners (MPII, NUA, UMIT, UniDU, MUNI, ETHZ)

- complementary expertise and potential for synergies
- collaboration started (dedicated 2-day workshop, bilateral visits)