

*27th Annual International ACM SIGIR Conference
- Workshop on Peer-to-Peer Information Retrieval -
Sheffield July 29*

Bookmark-driven Query Routing in Peer-to-Peer Web Search

**Matthias Bender, Sebastian Michel,
Gerhard Weikum, Christian Zimmer**

Max-Planck-Institute for Computer Science
Saarbrücken, Germany



MAX-PLANCK-GESELLSCHAFT

smichel@mpi-sb.mpg.de



Overview

- Motivation
- Peer-to-Peer Systems
- Related Work
- Design Fundamentals
- Bookmark-driven Query Routing
- Conclusion/Summary
- Ongoing and Future Work

Motivation

- “Why ask one if you can ask thousands?”
- Break information monopolies.
- Intellectual input from a large number of users.
- Use **bookmarks** to find relevant peers.

→ **Peer-to-Peer Web Search**

P2P Systems

- peer:
 - **“one that is of equal standing with another”**
 - *“one belonging to the same societal group especially based on age, grade, or status ”*

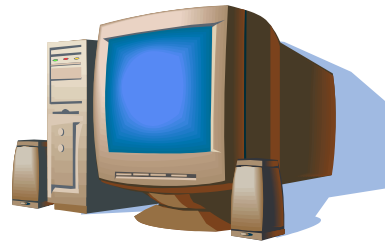
source: Merriam-Webster Online Dictionary

- Benefits
 - no single point of failure
 - resource/data sharing
- Problems/Challenges:
 - authority/trust/incentives
 - high dynamics
 - ...

Structured P2P-Systems

- Distributed Hashtable (DHT)
Highly efficient support of one “simple” method

lookup(key)



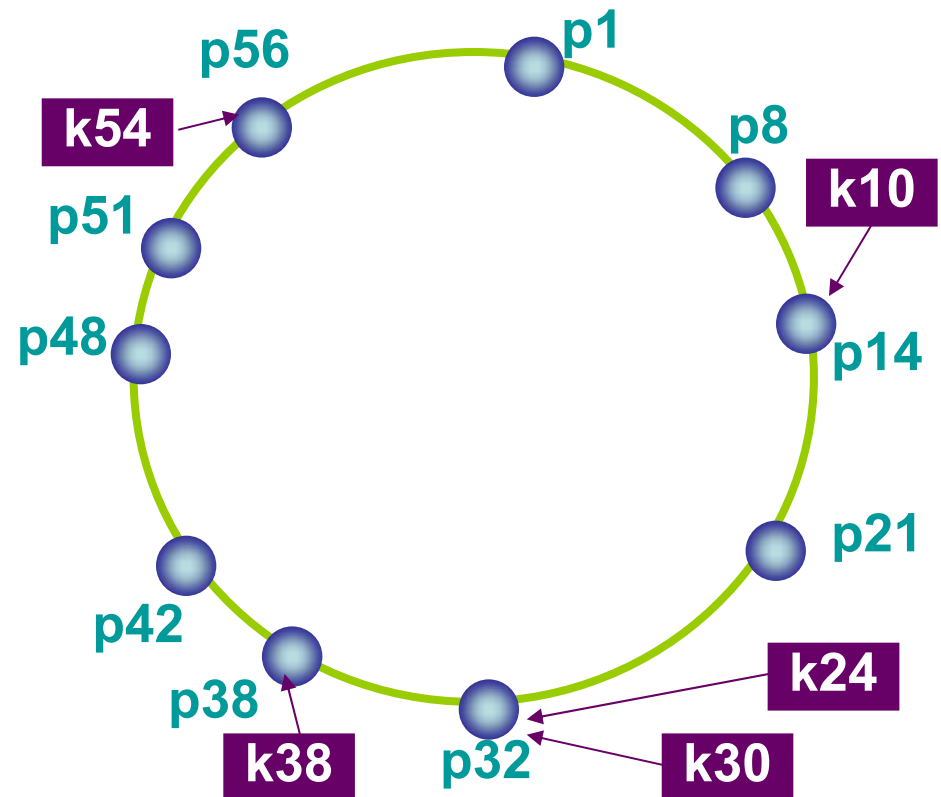
in $O(\log n)$ routing hops!

**+ robustness to
load skew,
failures,
dynamics**

- Chord: I. Stoica et al.
- CAN: S. Ratnasamy et al.
- P-Grid: K. Aberer

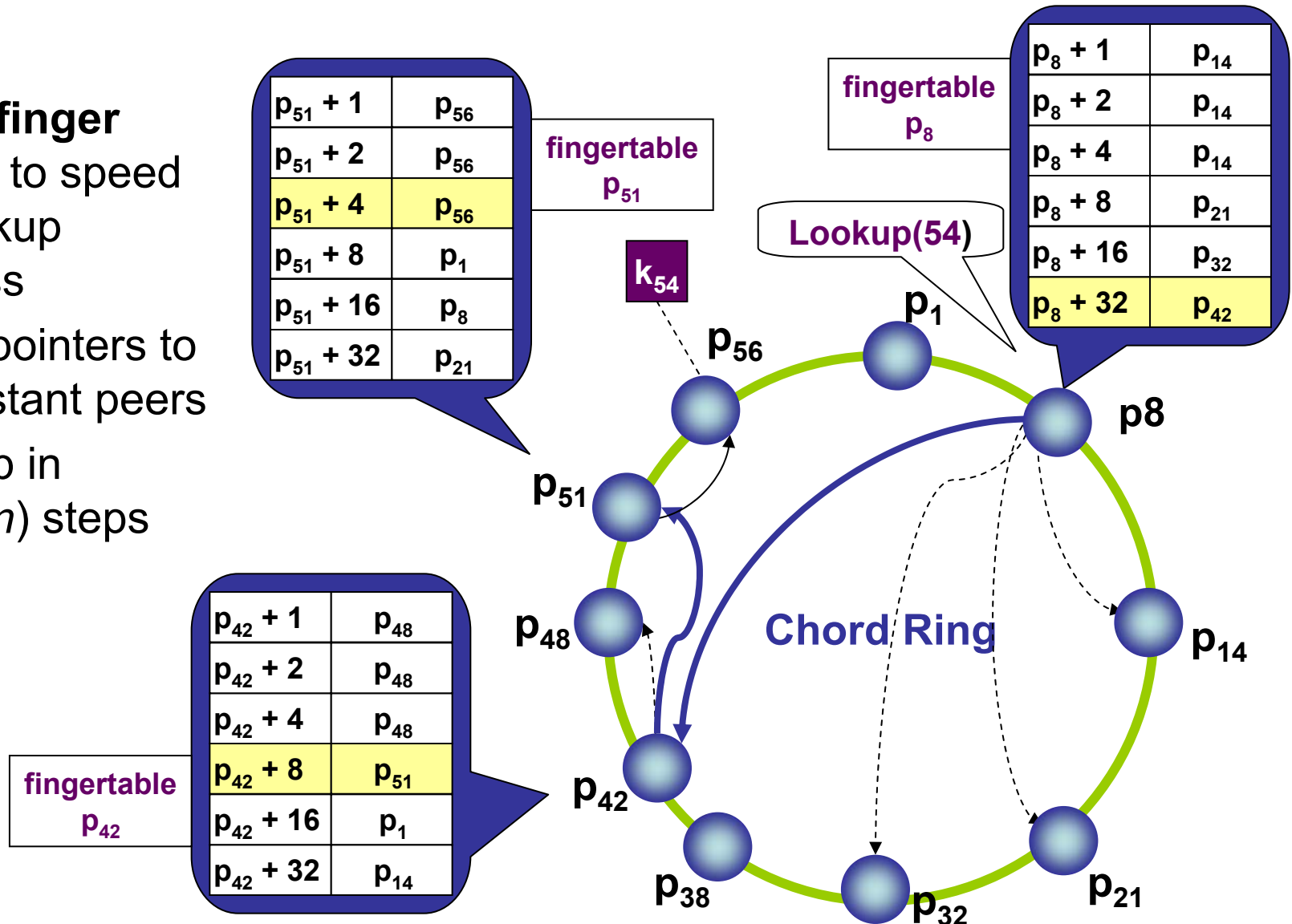
Chord

- Peers and keys are mapped to the same cyclic ID space using a hash function
- Key k (e.g., *hash(file name)*) is assigned to the node with key p (e.g., *hash(IP address)*) such that $k \leq p$ and there is no node p' with $k \leq p'$ and $p' < p$



Chord

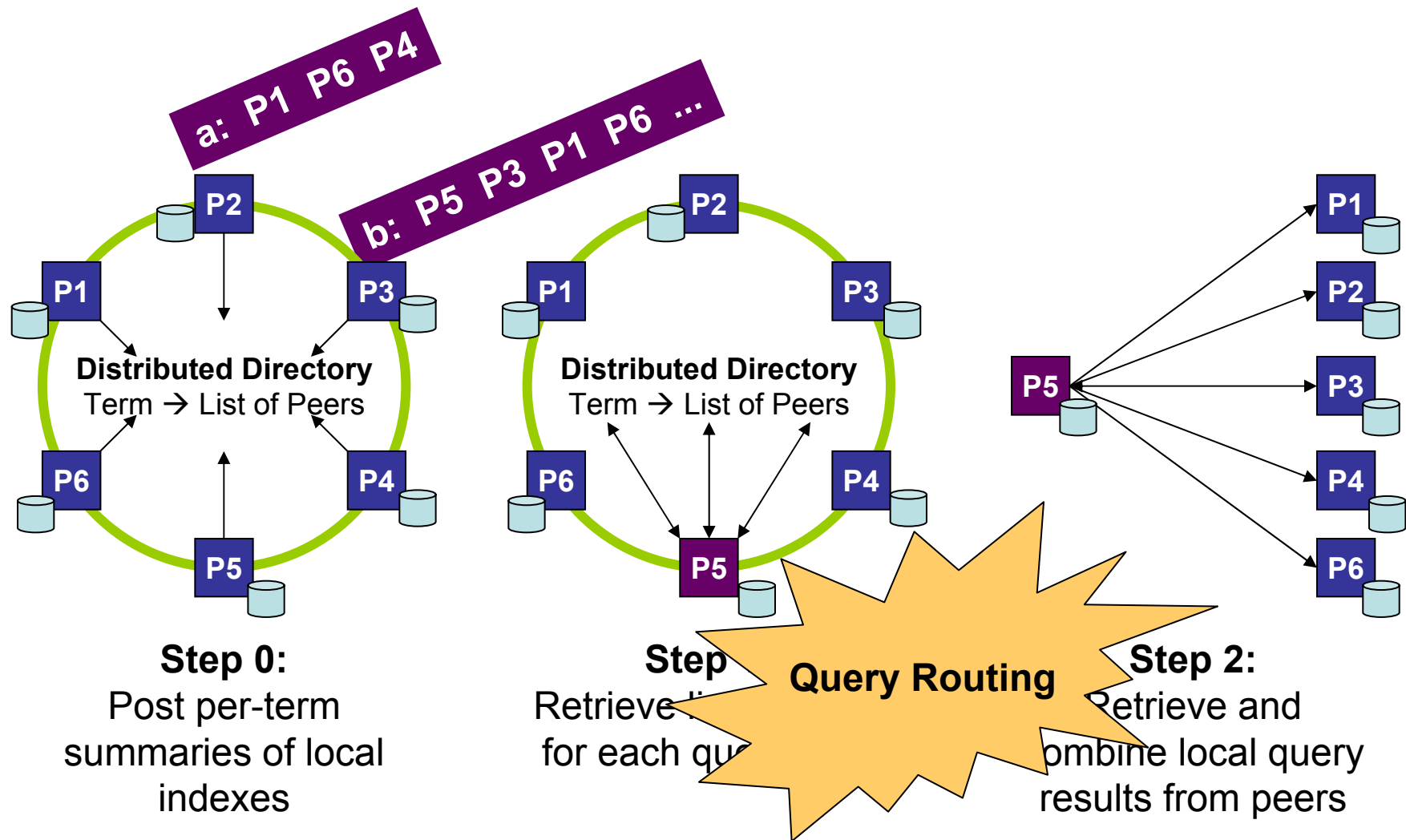
- Using **finger tables** to speed up lookup process
- Store pointers to few distant peers
- Lookup in $O(\log n)$ steps



Related Work

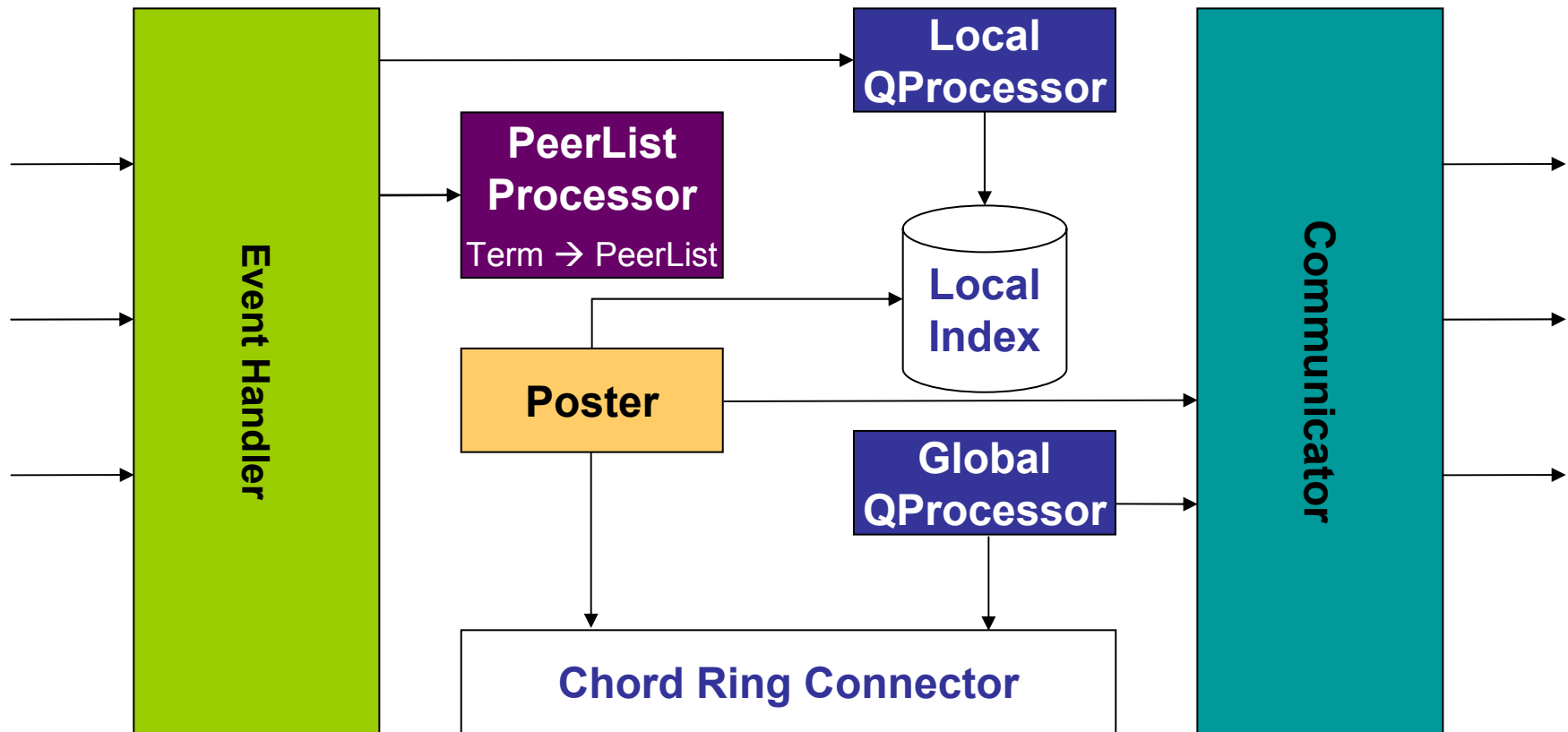
- Distributed IR:
 - CORI (*Callan et al., 1995*)
 - GLOSS (*Gravano et al., 1999*)
 - “A decision-theoretic approach to db selection in networked IR” (*Fuhr 1999*)
- P2P Search:
 - GALANX (*DeWitt et al., 2003*)
 - Odissea (*Suel et al., 2003*)
 - PlanetP (*Cuenca-Acuna et al., 2002*)
- Metasearch Engines

Design Fundamentals

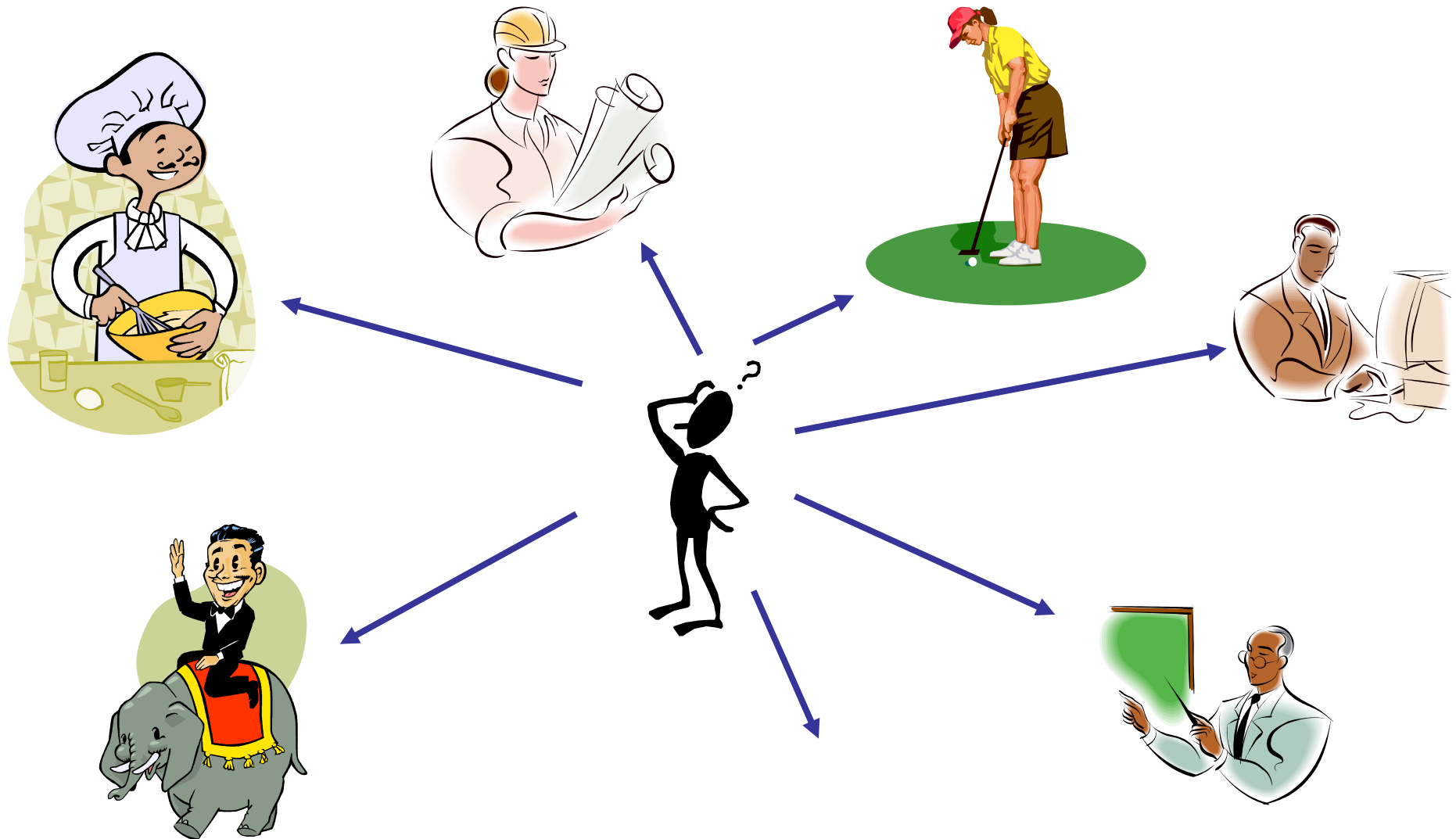


System Architecture

Architecture of a single peer



Why bookmark driven QR?



Bookmark-driven Query Routing

- bookmarks reflect user interests

bookmarks you have

- “Tell me what ~~books you read~~
and I tell you who you are”

 use bookmarks to find “relevant” peers

Relevance

- Notion of “relevant”

- **Similar content**
- **No or small overlap**

Bookmark-driven Query Routing

- Similarity between two peers?
 - Compare bookmarks instead of comparing local indexes (too expensive).
 - Assumption: Index has been created by focused crawling using bookmarks as crawl seeds.
- We can compare bookmark lists by
 - comparing the URLs
 - comparing term distributions of the documents referenced by the URLs

Information Similarity

- Kullback-Leibler distance (relative entropy)

$$KL(f, g) = \sum_x f(x) * \log\left(\frac{f(x)}{g(x)}\right)$$

where f and g are probability distributions.

- Measure for information inequality.

Overlap and Benefit

$overlap(A, B) := ca$

**this
is query independent:
can be precomputed, cached,...**

Query based peer assessment

- Calculate similarity between the query and the bookmarks.
- Use term distribution of the top- k local results

$$\textit{benefit}(B) \sim \frac{1}{KL(Q, B)} * \frac{1}{\textit{overlap}(A, B)}$$

Conclusion/Summary

- P2P approach for collaborative search
- Scalable Search engine
- Extensible system architecture
- Bookmark-driven query routing

Ongoing and Future Work

- Complete the implementation
- Experiments on real web data
- Replication

Prototype

P2P Search

Chord Component

Local Chord Port: 9001
 Local Application Port: 9002
 Remote IP: localhost
 Remote Chord Port:

name	value
chord id	36155
ip	139.19.54.20
ring exponent	65536
ring exponent	16
succ id	36155

Posts

d (59093)
 a (63874)
 c (38842)
 b (18590)

Queries

IP:Port	URL	Title	Score
Peer 127.0.0.1:9002	DOC 1	DOC 1	1.0986132886686...
Peer 127.0.0.1:9002	DOC 2	DOC 2	0.4785563631972...

Thanks for your attention.

Questions?

