Assisted Peer-to-Peer Search with Partial Indexing

Rongmei Zhang and Y. Charlie Hu

Presented by Patrick Roos

Aim

- improve search in unstructured peer-to-peer (P2P) overlay networks by building a partial index of shared data
- index maintains two types of information:
  - the top interests of peers
  - globally unpopular data

  both characterized by data properties.

Background

- Centralized P2P
  - Napster
  - Maintain a full index of shared files
  - Search is performed on centralized server
- Unstructured P2P
  - Gnutella
  - Flooding based search mechanism
- Structured P2P
  - Chord, Pastry, CAN
  - DHT based search mechanism

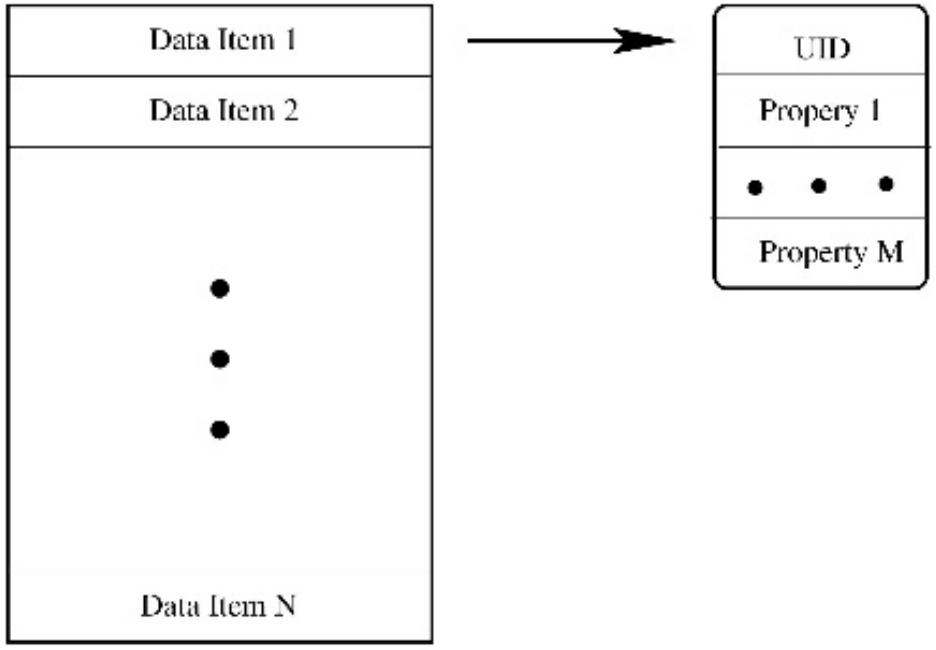
Assisted P2P Search Proposed

- DHT-based partial indexing scheme
- Instead of indexing all data items owned by each peer, only a portion of them are registered with the index
- Assisted search with partial indexing
  - Exploiting partial index to improve search efficiency in unstructured P2P overlay network

Properties

Data possessed by a peer. Each data item contains one or more properties

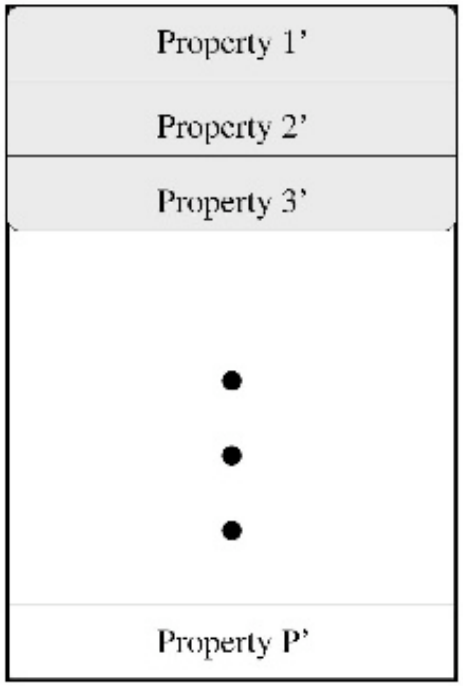
- Media metadata, e.g.: title, singer, genre
- Documents: top ranking keywords



Interests

Most frequent properties contained in the queries that it issues.

- Dominant properties of its data possession

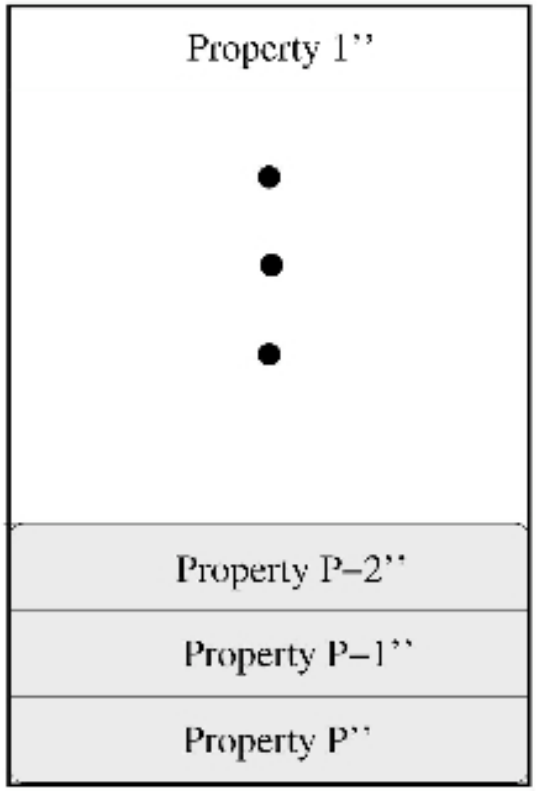


Popularity of Data

- with respect to all peers in the search network
- defined by its availability, i.e., the more popular the data, the more replicas are available and the easier it is to locate a replica
- frequency at which the data item is queried is a good estimation of its popularity
- Since queries are specified by properties of the data being searched for, we consider the *popularities of properties*

Popularity of Data

- Observed by individual peers
- Those properties of local data that are seen the least frequently in passing queries are identified as “unpopular
- Figure: Merged properties of local data items, sorted according to the frequency of appearance in received queries

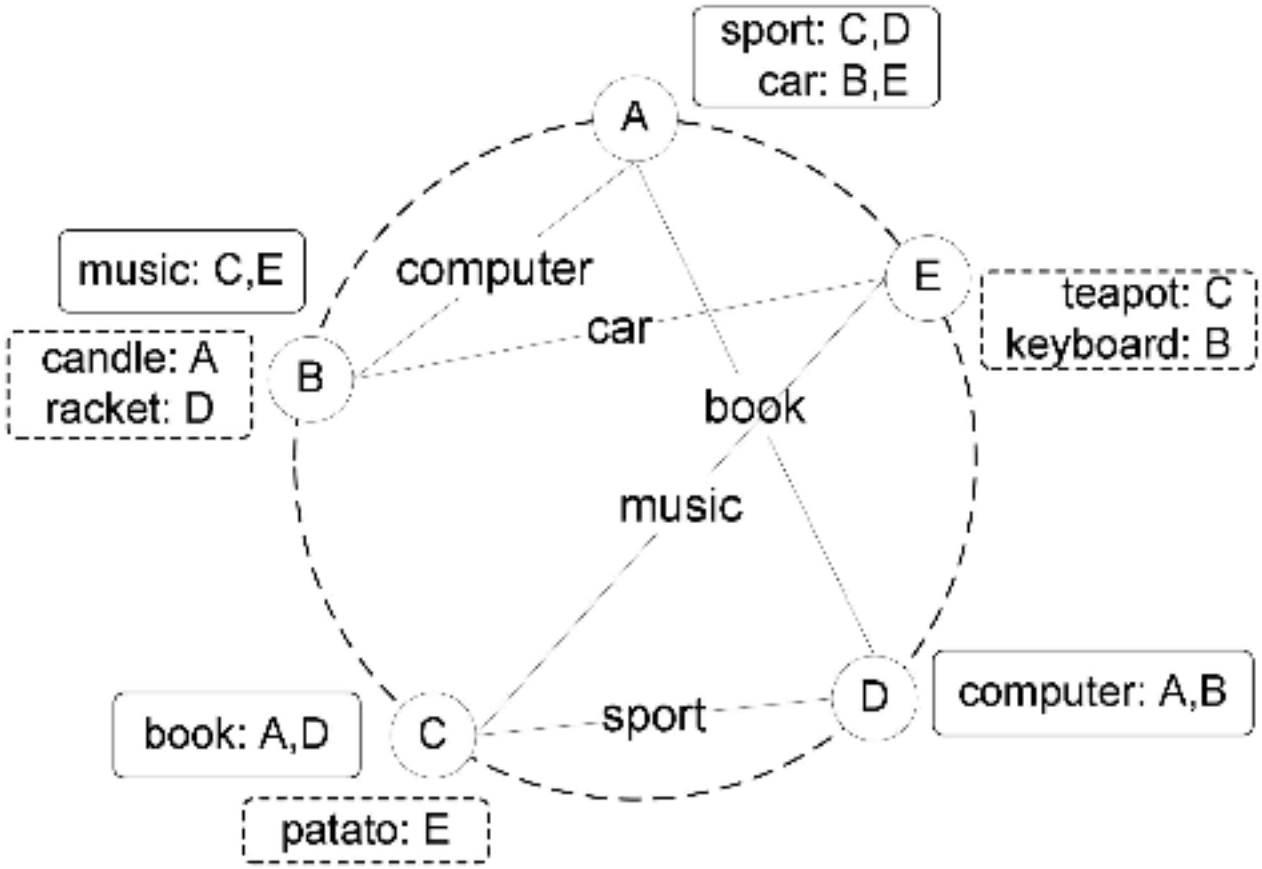


Assisted search with partial indexing

Two Logical Overlays

- Index overlay
  - Structured P2P network
  - Pastry
- Search overlay
  - Unstructured P2P network
  - Gnutella

Assisted Search Network



Index overlay

- The index maintains two types of information:
  - the top interests of peers
    - The interests of a peer are represented by the dominant properties of its data possession
  - globally unpopular data
    - Property popularity can be determined from observing passing traffic
    - Properties of local data that are seen the least frequently in passing queries are identified as "unpopular

Purpose of Index

- The index overlay assists the search overlay to improve its search performance in the following three way
  1. assists peers to find other peers with similar interests
  2. provides search hints for those data difficult to locate
  3. improve the chances of finding unpopular data

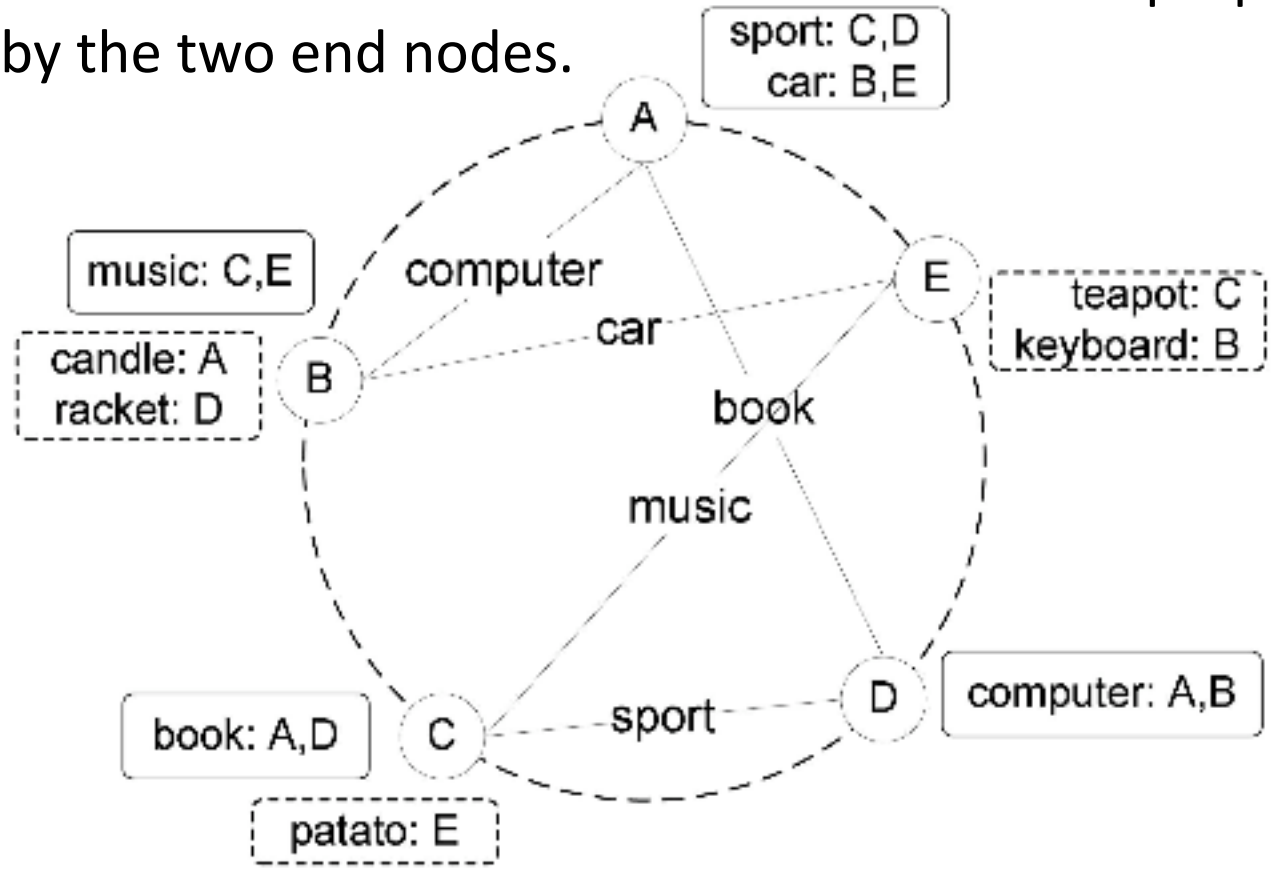
Assisting peers to find other peers with similar interests

- peers communicate their interests via the index overlay and the search overlay is constructed based on peer interests
- each peer registers its own major interests with the index overlay and also looks up other existing peers sharing the same interests
- In this way, a peer always connects to those peers in the search overlay that share common interests

Interests

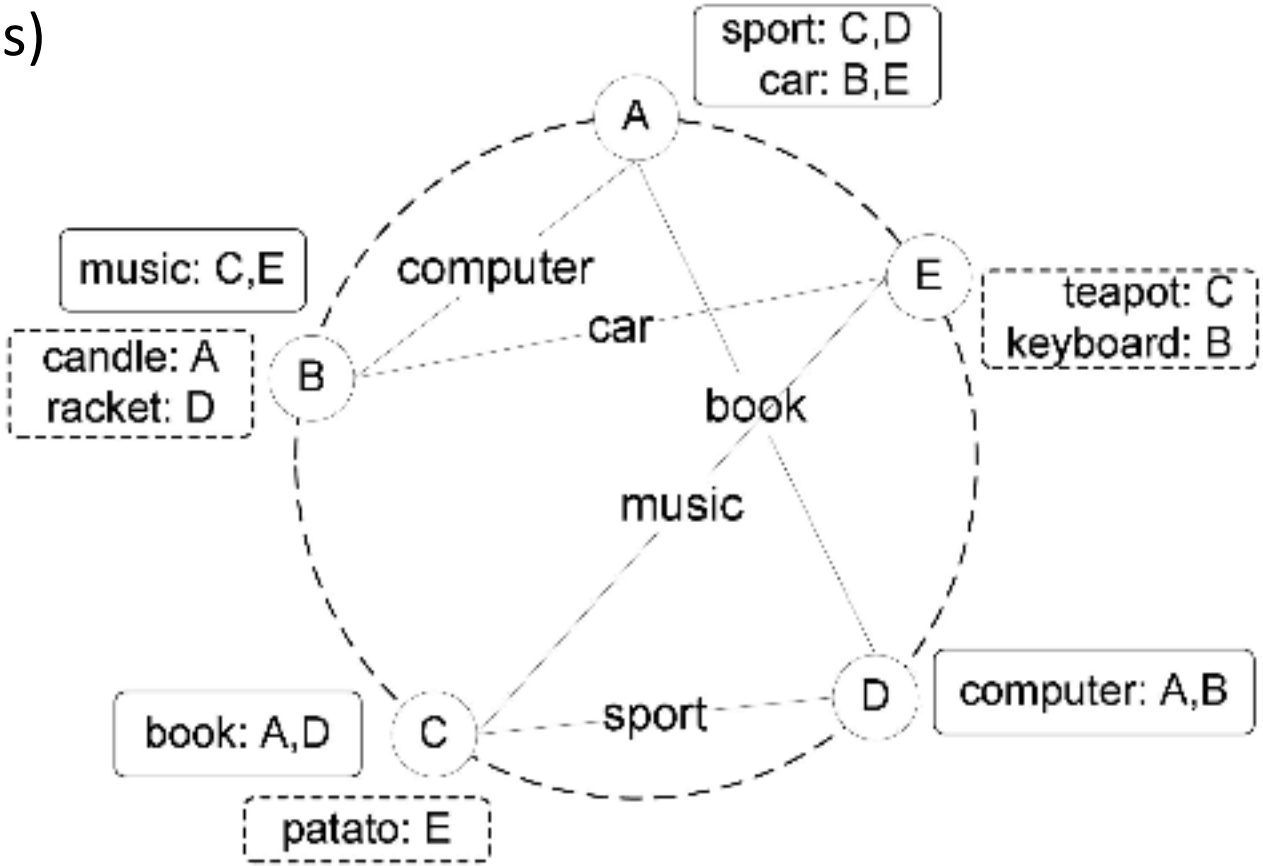
• Each of nodes A, B, C, and D maintains registrations for at least one property (solid boxes)

• Each node is connected to two other nodes in the search overlay, and each link is labeled with the associated interest property shared by the two end nodes.



Globally Unpopular Properties

- Are explicitly registered in index overlay
- Can be returned as potential destinations when the index overlay is queried for search hints (together with interest-based registries)



Search Overlay

- A Search overlay is created based on peer interests
- After joining the index overlay, a new peer can obtain the addresses of other peers with similar interests
- As the reply containing such nodes is received, the peer initiates connections to them

Search

- A query is first issued to the search overlay
- (e.g. controlled flooding, random walk)
- If the first try in the search overlay yields no hits at all or the peer is not satisfied with the results
- The peer has a second chance by seeking search guidance from the index overlay

Experiments

- Simulations based on traces from a real Gnutella system (Gtk-Gnutella)
- record Query and QueryHit messages
- 1,703 peers, 518,909 files, and 129,363 words
- about 57 % keywords are seen in only one filename and more than 80 percent in at most five filenames
- large number of keywords are not likely to be selected as peer interests

Experiments

- about 87 % of keywords are accessed by five or fewer peers
- popularity of keywords is highly skewed in the Gnutella trace
- Before the simulation starts, each file is placed at the node from where it is first seen in a QueryHit message
- Since each Query message may yield more than one matching results, the querier randomly selects one file for downloading

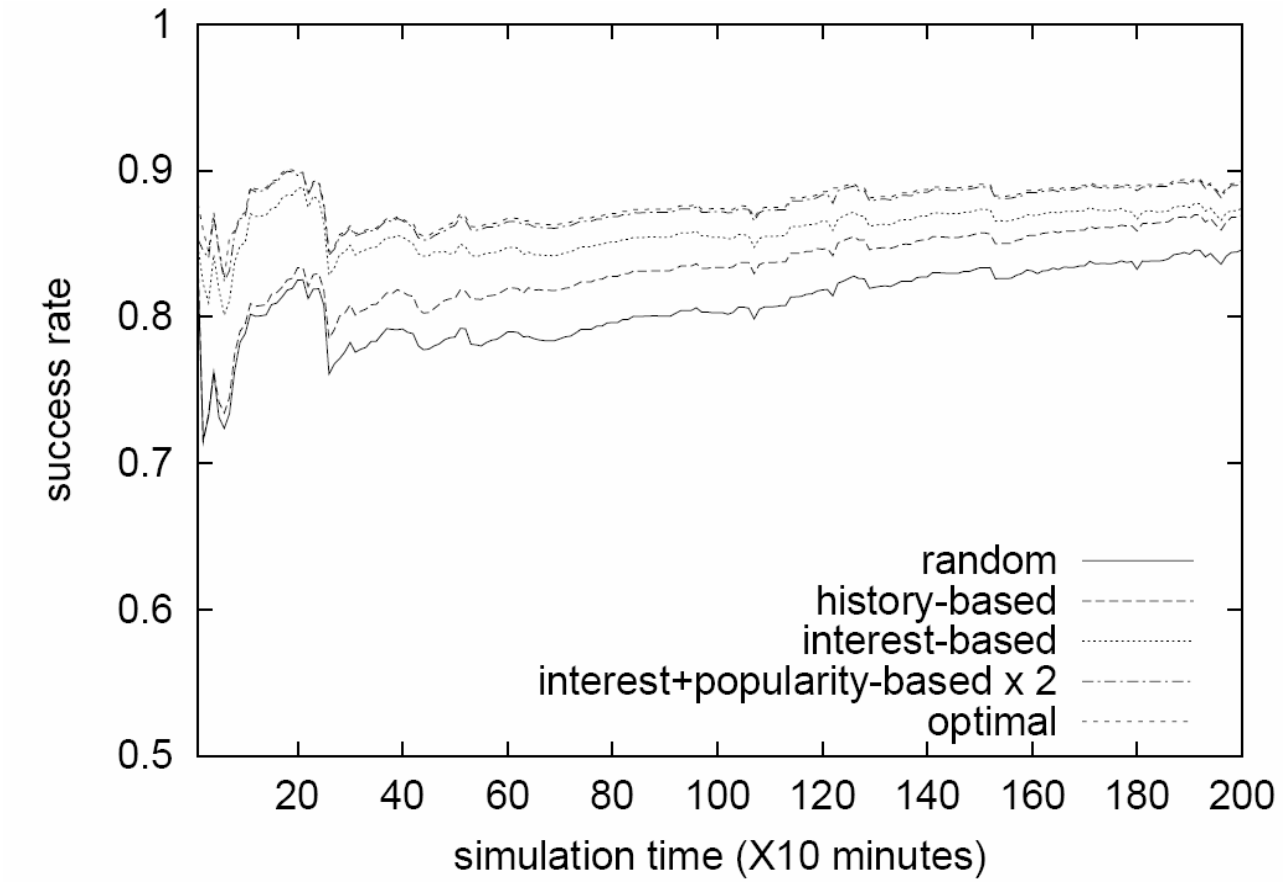
Experiment Protocol Config

- each peer in the trace participates in both the index overlay and the search overlay
- The index overlay is a Pastry (FreePastry) network
- In the search overlay, each peer is connected to four neighbors

Compared Against:

- Gnutella protocol (labeled “Gnutella”), where the search overlay is formed randomly and flooding is used for resolving queries

Success Rate



Performance

Search Delay

- Without second-chance query hint from index overlay: same
- With second chance: ~ 2.5 hops higher

Avg Bandwidth consumption at each Node

- About 20 – 30 Bps more

Conclusion

- The assisted search protocol achieves higher search efficiency and scalability than a pure flooding-based or history- based search scheme
- At the same time it also retains desirable features of search in unstructured overlays such as robustness