

IQN Routing: Integrating Quality and Novelty for P2P Web Search

EDBT 2006, Munich, Germany

Sebastian Michel ^{*}, Matthias Bender ^{*}, Peter Triantafillou [◇],
Gerhard Weikum ^{*}

^{*} Max-Planck-Institut für Informatik
66123 Saarbrücken, Germany

[◇] RACTI and University of Patras
26500 Rio, Greece

27th March 2006

Outline

- 1 Introduction
- 2 Distributed Web Search with Minerva
- 3 IQN Routing
- 4 Evaluation
- 5 Conclusion and Outlook

P2P Systems

Became famous through file-sharing applications, like Gnutella, KaZAA, Napster.

Applications like:

- Internet telephony (e.g. Skype)
- Filesharing
- Pub/Sub

Question:

Is there an interesting and legal P2P application?

Motivation

Why P2P Web Search?

- Benefit from intellectual input from a large user community. (Bookmarks, click-streams, ...)
- Break information monopolies
- Coverage of the web
- Exploit mostly idle resources

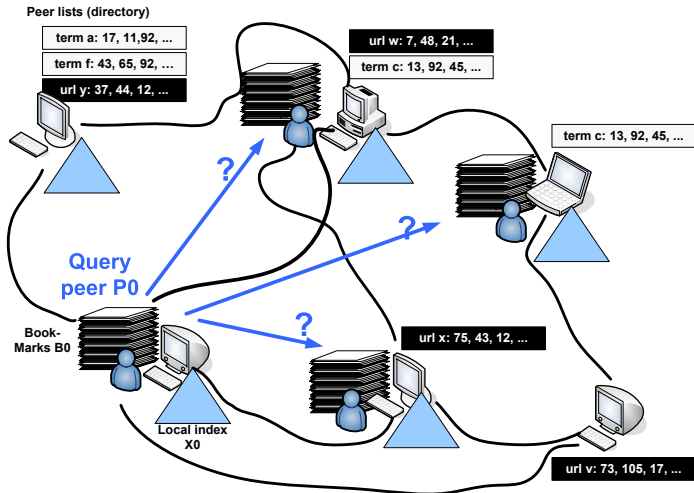
Related to distributed IR, but some additional aspects

- high dynamics
- each peer has its own collection
- peers are independently crawling the web

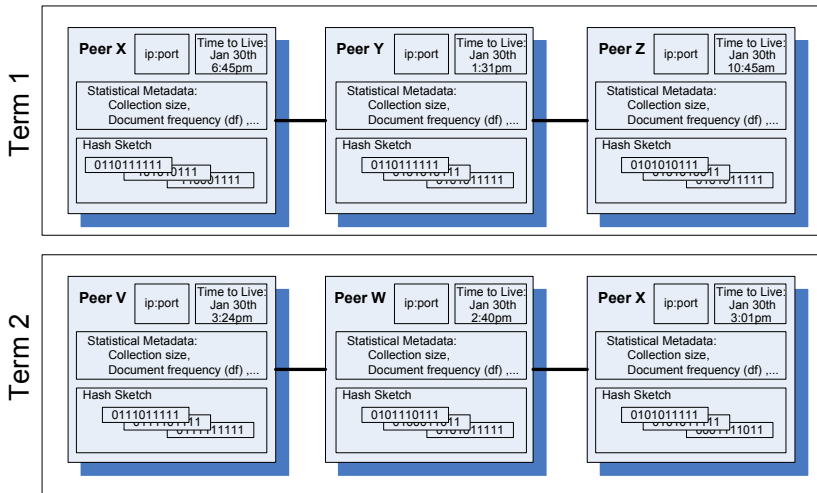
Minerva Design Fundamentals

- Peers with local collections, e.g., built by focused crawler. Tailored to the users' specific interest profiles.
- Peers share metadata about local indexes
- Form physically distributed *term* → *peer* directory
- Layered on top of DHT
- Peers use directory to discover promising peers for query

Minerva System Architecture



Inside Peerlists ...



How to find promising Peers?

State of the art: Find peers with high quality documents.

Existing Strategies:

- based on per-term metadata
- combine metadata for all query terms
- select peers with highest expected result quality

Why Quality-Only is not Enough!

Problem:

Peers crawl the web independently.

→ **overlapping collections**

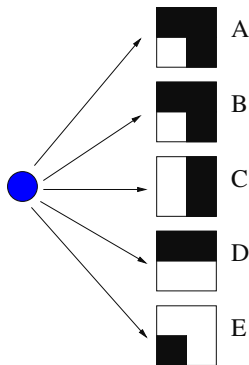
(For instance, cnn.com might have been crawled by a lot of peers)

Goal: Integrate Quality and Novelty (IQN).

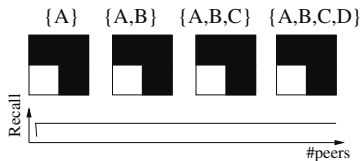
→ achieve high recall with fewer peers than traditional approaches

■ **We define:** $Usefulness := quality * novelty$

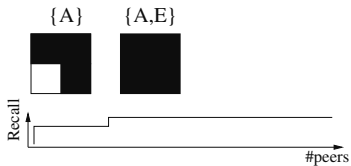
Example



naïve routing strategy



overlap aware routing strategy



IQN Algorithm: Integrating Quality and Novelty

Extend per-peer, per-term meta-data with synopses that describe local collections.

Select all peers to query a-priori

Based on statistics (not their actual query results).

- Choose first Peer X based on quality only → use X's per-query descriptor as initial representation of already seen documents.
- Then choose Peer Y with the highest usefulness w.r.t. the already seen docs.
- Merge representations for the peers selected so far and iterate.

Quality Prediction

Find best suitable peers based only on quality measures.

Most popular selection strategies:

- CORI (Callan et al.)
- GIOSS (Gravano et al.)
- Decision-theoretic framework (Fuhr)

Based on:

- document frequency
- maximum term frequency
- total number of distinct terms
- total number of documents

Novelty Prediction

Add statistics that allow novelty estimation.

We are interested in $|S_B - (S_A \cap S_B)|$

Two possible approaches:

- represent whole collection
- use separate representations for (term-specific) index lists

Term-specific representations allow query-specific overlap estimation!

Multi-keyword queries

Combine per-term descriptors of a peer to form per-query descriptor

Data Synopses

Searching for appropriate data synopses ...

Requirements:

- Compact
- Highly accurate
- (.....)

- Bloom Filter
- Hash-Sketches
- Min-wise independent permutations

Bloom Filter

Bit-array of length m . Insert documents by settings bits using k hash functions.

Membership-Queries: A document is contained in a Bloom filter if the corresponding are set. Problem: false positives

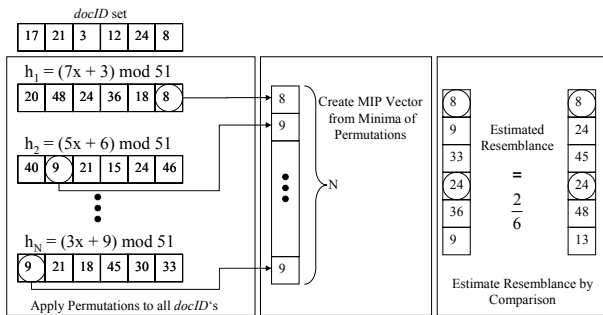
$$pfp \approx (1 - e^{-kn/m})^k$$

Hash Sketches

- Pseudo-uniform hash function h
- Apply h to all documents and record the position of the least significant (leftmost) 1-bit in the binary representation in a bitmap vector $B[0 \dots L - 1]$.
- Idea: $B[0]$ will be set approximately $\frac{n}{2}$ times, $B[1]$ approximately $\frac{n}{4}$ times,
- More formally: The leftmost 1 bit at position k provides an estimation of $\log(n)$.

Min-wise independent Permutations (MIPS)

N independent permutations



Given $|S_A|$, $|S_B|$, and the resemblance $R = \frac{|S_A \cap S_B|}{|S_A \cup S_B|}$, we estimate the overlap between S_A and S_B as $|S_A \cap S_B| = \frac{R * (|S_A| + |S_B|)}{(R+1)}$.

Aggregate Synopses

Remember: After having selected the best peer in an iteration of the IQN method, we need to update the reference synopsis.

MIPs

Given $MIP_{SA}[]$ and $MIP_{SB}[]$, one can form $MIP_{SA \cup B}[]$ as follows
$$MIP_{SA \cup B}[i] = \min\{MIP_{SA}[i], MIP_{SB}[i]\} \quad \forall i : 1 \leq i \leq n.$$

Hash Sketches

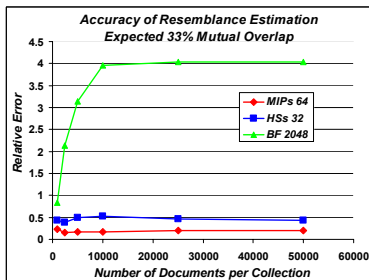
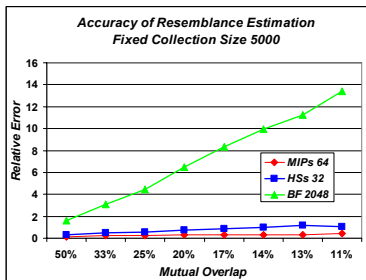
straight forward: use the bit-wise *OR* operation

Bloom Filter

Again, bit-wise *OR*: $BF_{A \cup B}[i] = BF_A[i] \text{ OR } BF_B[i] \quad \forall i : 1 \leq i \leq n.$

Accuracy Evaluation

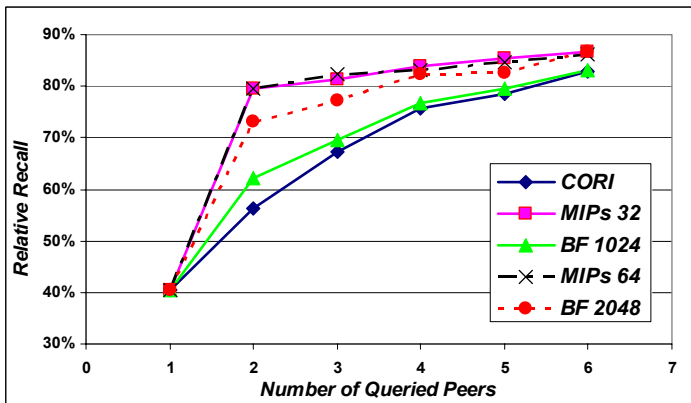
Experiments on synthetic data-sets:



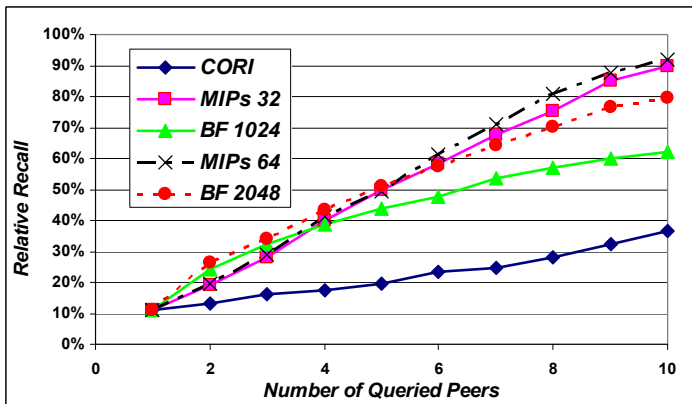
Experimental Setup (2)

- Routing Strategies
 - pure CORI
 - overlap aware CORI using MIP based overlap prediction
 - overlap aware CORI using Bloom Filter based overlap prediction
- Datasets:
 - Subset of the official TREC .GOV collection split into disjoint fragments. Building peers using
 - sliding window over these fragments
 - mirrored collections
 - ...
 - Queries: 50 TREC-2003 Web queries, e.g. juvenile delinquency
 - Measure the recall w.r.t. the query results of the whole document set (relative recall)

Experimental Results: $\binom{6}{3}$ Benchmark



Experimental Results: Sliding Window Benchmark



Conclusion and Outlook

Conclusion

- Comprehensive performance evaluation of MIPs, Bloom Filter, and Hash-Sketches
- Experiments on real-web data showing the impact of overlap aware query routing

Future Work

- Evaluation of the usage of histogram enhanced synopses
- Adaptive synopses lengths
- System behavior under churn