Global Document Frequency Estimation in Peer-to-Peer Web Search

WebDB 2006

Matthias Bender *, Sebastian Michel *, Peter Triantafillou ○, Gerhard Weikum *

* Max-Planck-Institut für Informatik ○ RACTI and University of Patras

June 30, 2006
Outline

1. Introduction
2. Distributed Web Search with Minerva
3. Global DF Estimation
4. Evaluation
5. Conclusion and Outlook
Motivation

Peer-to-Peer
- Became famous through file-sharing applications like Gnutella, KaZAA, Napster
- Today: Applications like: Skype, pub/sub, Web search

Why P2P Web Search?
- Benefit from social networks for more powerful IR models
- Break information monopolies
- Exploit mostly idle resources

Related to distributed IR, but some additional aspects
- High dynamics
- Overlapping collections from autonomous peers
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 \textit{term} $\rightarrow$ \textit{peer} directory
- Layered on top of DHT
- Peers use directory to discover promising peers for query
Scoring in IR

Usually weighted sum

\[ s(d) := \sum_{t \in Q} \frac{1}{DF(t)} \times TF(t, d) \]

where \( DF = \) Document Frequency, and \( TF = \) Term frequency.
Problem Statement

Lack of Global Statistics

- No global DF values available, peers use local DF → document scores incompatible
- “Good” peers with many documents have high DF → low local scores → documents from bad peers boosted

Goal: Estimate global DF in the presence of

- overlapping collections (global DF ≠ sum of local DF’s!!!)
- network dynamics

without additional messages
→ Scores compatible, result merging trivial
Global Counting

**Example 1**

How many distinct movies are available in the P2P system?
- high degree of replication (current top movies replicated probably a few hundred times)
- no global knowledge (no central manager like in Napster)

**Example 2**

Counting the number of persons at SIGMOD ’06:
- everybody participates in counting
- cannot take max: nobody has seen all participants
- high overlap: summing up is not accurate
Centralized setting: Hash sketches as multiset cardinality estimator.

- 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 $B[0 \ldots L - 1]$.
- Idea: $B[0]$ will be set approximately $\frac{n}{2}$ times, $B[1]$ approximately $\frac{n}{4}$ times, ..... 
- More formally: The rightmost 1 bit at position provides an estimation of $\log(n)$.
- Use multiple bitmap vectors to increase accuracy.
Hash Sketch Basics

Hash Sketches

Distributivity Theorem:

Let \( \beta(S) \) be the set of bit positions \( \rho(h(d)) \) for all \( d \in S \), and \( \rho(y) = \min_{k \geq 0} \text{bit}(y, k) \neq 0, \ y > 0. \)
Then \( \beta(S_1 \cup S_2) = \beta(S_1) \cup \beta(S_2). \)

Example

Let \( hs_A \) be the hash sketch describing set \( A \), and \( hs_B \) the hash sketch of set \( B \).
Then \( hs_A \ OR_{bit-wise} hs_B = hs_{A \cup B}. \)
Directory based DF Estimation

Directory Maintenance

- Include per-term hash sketch in term-specific post
  → No additional messages
- Retrieval combined hash sketch while retrieving Peerlists
  → No additional messages

Usage in Query Execution

- Send estimated DF as weights to queried peers
- Local QE, reweight on-the-fly with global DF
Evaluation

Experiment 1
General accuracy of hash-sketch based cardinality estimation. 
*omitted here for time reasons*

Experiment 2
Accuracy of global df estimation in dynamic networks

Experiment 3
Impact of global df in P2P search
Experiment 2: Accuracy of DF Estimation under Churn

Data/Peers

- synthetic data (document ids)
- 256 bitmap hash-sketches
- Initially 1000 Peers + Entering/Leaving Peers
Experiment 2: Accuracy of DF Estimation in dynamic Networks

256 bitmap hashsketch, synthetic collections
Experiment 3: Impact on Result Quality

Dataset
- 10 thematic web collections, split into 4 fragments each
- Created 40 peers by creating all 3-subsets for each topic

Queries
- 30 popular Google queries (Zeitgeist)

Quality Measure
Distance of
- global DF based merging
- local DF based merging
- CORI [Callan et al.] based result merging (normalization)
  to hypothetical centralized ranking
Experiment 3: Impact on Result Quality

40 peers

![Graph showing the impact on result quality for different methods.](image-url)
Conclusion and Outlook

Conclusion
- New method for global df estimation in large scale P2P networks
- Experiments in dynamic networks
- Experiments on real-web data

Future Work
- Evaluation of the impact using relevance assessments