



Search in the COAST Project

B. Barla Cambazoglu
Yahoo! Research

Disclaimer

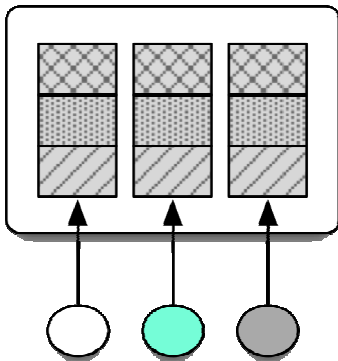
- This talk presents the opinions of the author. It does not necessarily reflect the views of Yahoo! Inc. or any other entity.
- Algorithms, techniques, features, etc. mentioned here might or might not be in use by Yahoo! or any other company.

Overview of COAST

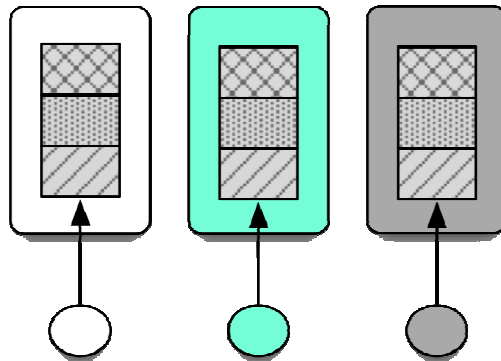
- Main concept
 - making the content accessible by content id rather than its location
- Our contribution
 - making the content in COAST searchable
- Auxiliary ideas
 - caching the content across the network
 - closest copy detection
 - obtaining content popularity from the network
 - discovering content passively

Large-Scale Search Engine Architectures

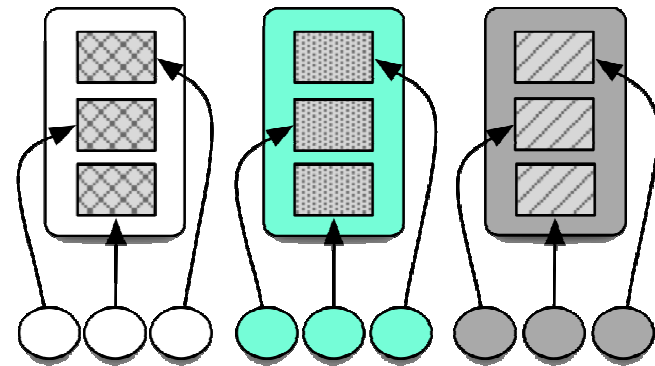
- centralized
 - 90s
 - not scalable



- replicated
 - current
 - not efficient

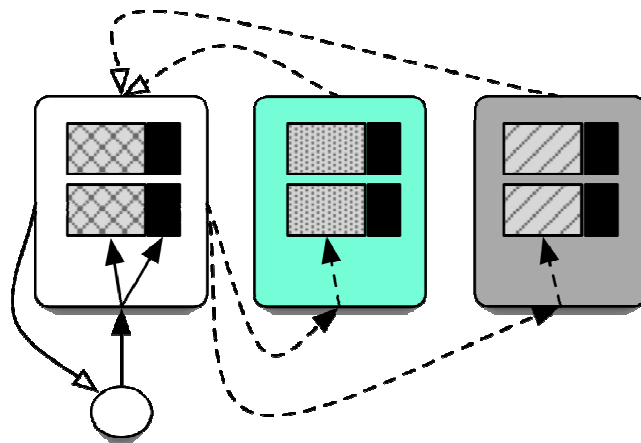
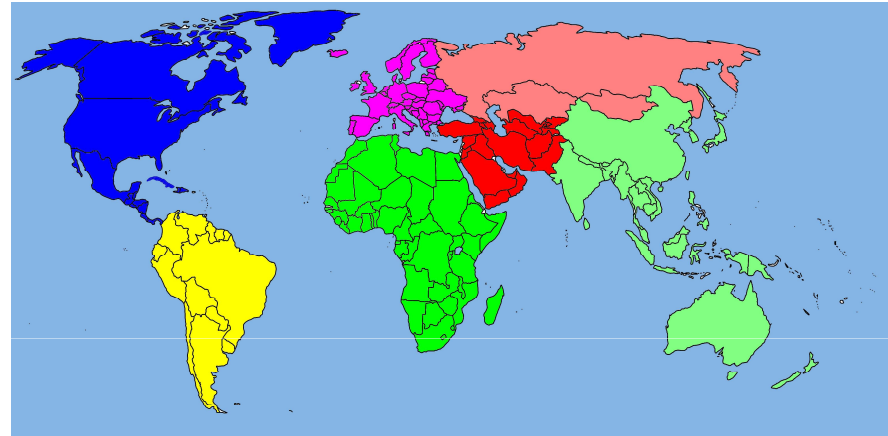


- partitioned
 - hypothetical
 - not effective



Multi-site Search Engine Architecture

- Features
 - geographically distant, local data centers
 - fully distributed crawling
 - partitioned inverted index
 - partial document replication
 - selective query forwarding



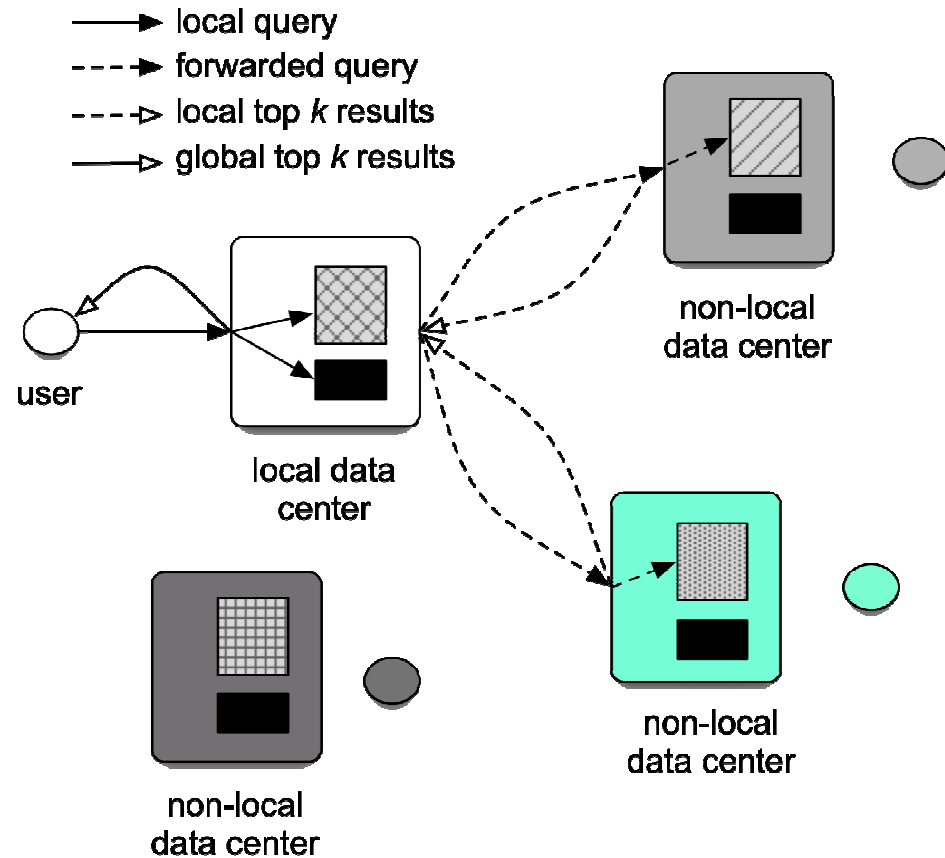
- Envisioned benefits
 - high crawling throughput
 - reduced query response time
 - reduced query workload
 - increased availability

Architectural Components

- Main components
 - query processing
 - threshold-based query forwarding
 - indexing
 - network feedback for content popularity
 - web crawling
 - passive URL discovery

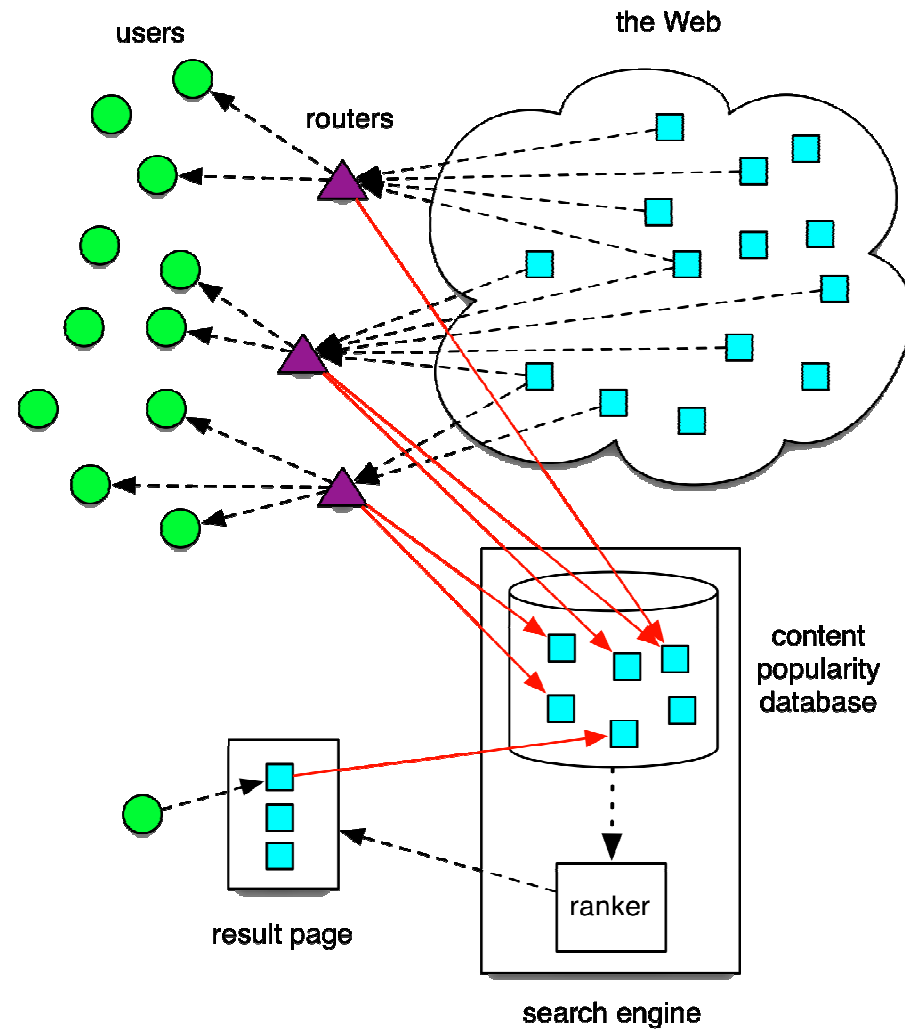
Query Processing

- Features
 - selective forwarding
 - false positives
 - increased workload
 - increased latency
 - false negatives
 - low result quality
 - our algorithms guarantee no false negatives



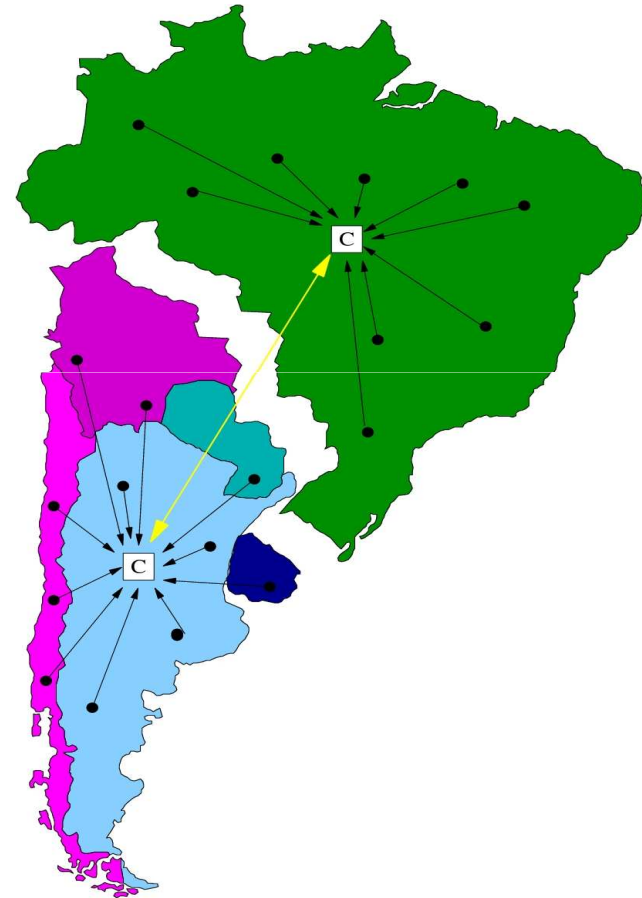
Indexing

- Traditional features
 - statistical analysis
 - link analysis
 - proximity
 - spam
 - clicks
 - session analysis
- A new feature
 - network feedback



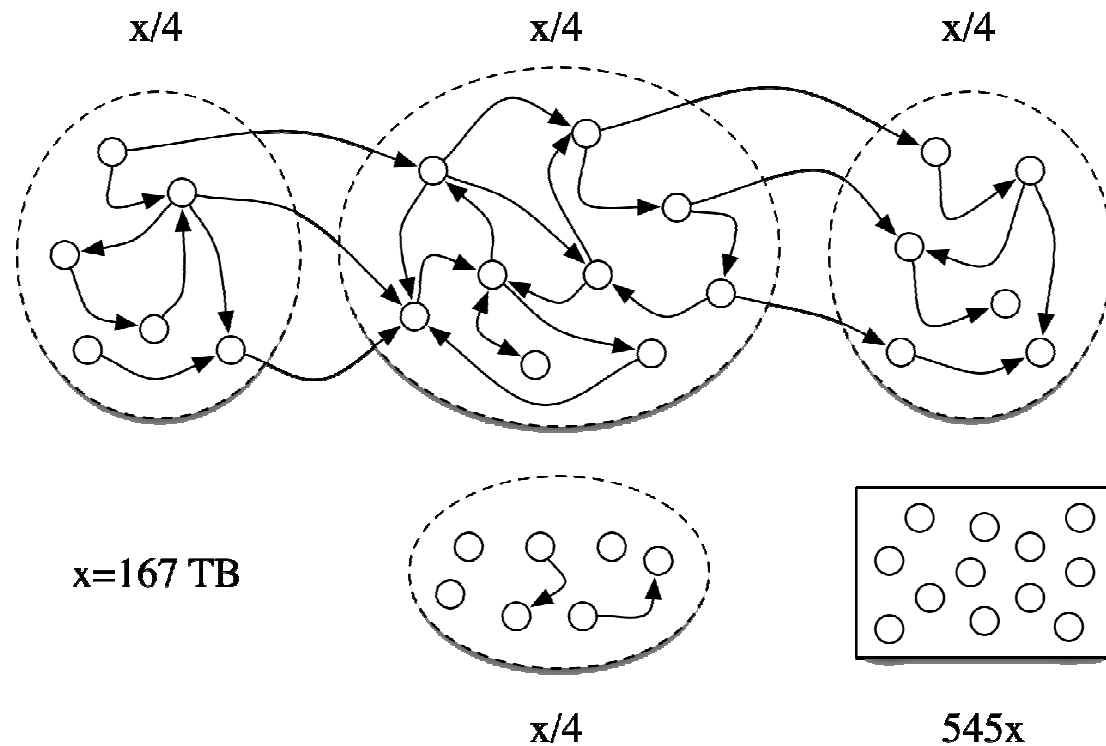
Web Crawling

- Higher crawling throughput
 - spatial locality
 - low latency
- Improved network politeness
 - less overhead on routers
- Resilience to network partitions
 - better coverage
- Increased availability
 - continuity of business
- Better coupling with distributed indexing/search
 - reduced data migration



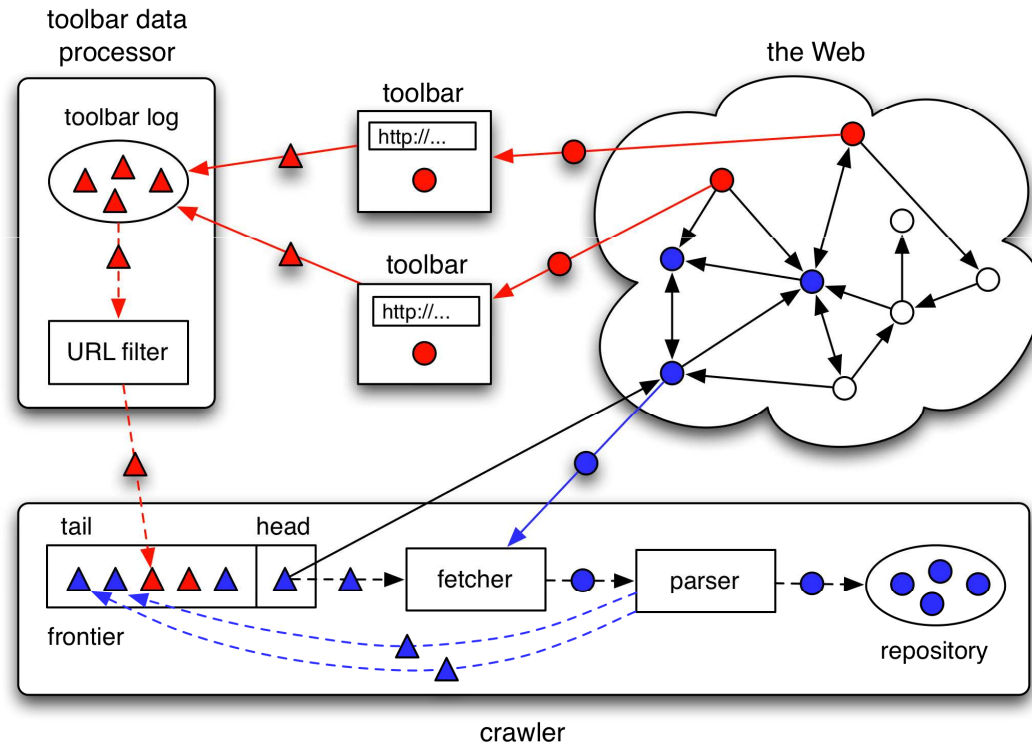
Web Crawling

- Challenges
 - tail content
 - hidden web
 - real-time content



Web Crawling

- Benefits
 - improved coverage
 - real-time discovery
 - reduced cost



People

- People
 - B. Barla Cambazoglu
 - Flavio Junqueira
 - Ivan Kelly
 - Vincent Leroy