Parallel Crawlers

Junghoo “John” Cho
University of California, LA

Hector Garcia-Molina
Stanford University
What is a Crawler?

- Initial URLs
- To Visit URLs
- Visited URLs
- Web Pages

1. Init
2. Get Next URL
3. Get Page
4. Extract URLs
Central vs Parallel
Parallel Crawler?
Why Study of A Parallel Crawler?

- Many advantages
  - Imperative for large-scale crawling
  - Can be run on cheaper machines
  - Network load dispersion
  - Network load reduction
- Hasn’t it been solved?
  - Little discussion in open literature
Outline

- Issues of parallel crawlers
  - Evaluation metrics
- Design alternatives
  - Parallel crawling models
  - Experimental evaluation
Issues?

• How much overhead?
  • Communication overhead?
  • Overlap?

• Will it be of same quality?
  • Page “importance”?
  • Web coverage?
Evaluation Metrics

- Communication overhead
  \[
  \frac{\text{No of exchanged messages}}{\text{No of page downloads}}
  \]

- Overlap
  \[
  1 - \frac{\text{No of unique pages downloaded}}{\text{No of page download by overall crawler}}
  \]

- Coverage
  \[
  \frac{\text{No of pages downloaded by the parallel crawler}}{\text{Total no of reachable pages}}
  \]
Evaluation Metrics (cont)

- Quality
  - An importance metric, say, backlink count
  - When we downloaded $k$ pages

\[
\frac{|\text{Download}_k \cap \text{Top}_k|}{|\text{Top}_k|}
\]

$\text{Top}_k$: top $k$ most important pages
$\text{Download}_k$: downloaded $k$ pages
Our Approach

- Identify design alternatives
- Compare them using real Web data
  - Result may be valid only for our dataset, but provides a good first look
- Mostly experimental study
  - Not much theoretical modeling and analysis
  - Theoretical study challenging due to lack of good Web model
- Future work
Experimental Dataset

- 40M pages
- December 1999 snapshot
- WebBase crawler
  - High indexing speed ~ 100 pages/sec
  - Large repository, currently ~ 120M pages
- Started from open directory pages
  Followed links in the breadth-first manner
Outline

- Issues of parallel crawlers
  - Evaluation metrics
- Design alternatives
  - Parallel crawling models
  - Experimental evaluation
Parallel Crawling Models

- Many different alternatives
  - Independent vs coordination?
  - Static partitioning vs dynamic assignment?
  - No communication vs URL exchange?
  - ...

- Briefly discussion on some of the issues
  - More details in the paper
Parallel Crawling Models

- Independent vs. Coordination?
Independent vs Coordination

- Independent
  - No communication
  - Major issue: Overlap? Coverage?
- Coordination
  - Major issue: communication overhead
- Experiments show significant overlap for independent model
  - E.g., Overlap = 2 for 90% coverage (8 processes)
Static vs Dynamic Coordination
Static vs Dynamic Coordination
Static vs Dynamic Coordination

- Dynamic coordination
  - More adaptive
  - Communication between crawlers and the coordinator may become bottleneck
  - May not be suitable to geographically-distributed crawlers

- Static assignment
  - Less adaptive
  - Less coordination overhead

- Focus on static assignment
Static Assignment

- How to partition the Web?
  - Site-based? URL-based? Domain-based?

- Do we need coordination?
  - Coverage issue: Can we discover all URLs?
  - Quality issue: Can we download “important” pages?
Coverage Issue

P1

P2

d, e cannot be reached
Coverage

No URL exchange. Starting from 5 random URLs
Quality Issue

- Crawling strategy
  - Estimate “importance” or “relevance” of pages as we crawl, and download important ones first
- Many importance metrics depend on link structure
- Need to how many pages in other partitions are pointing to a page
- Link exchange necessary
Communication Issue

- Important especially when crawlers are geographically distributed
- Techniques to discuss
  - Batching: send a batch of links periodically
  - Replication is also studied in the paper
Impact of Batching on Quality

Importance metric: Top 5M most-linked pages
Related Work

- Page selection
  - Focused crawling
- Page refresh
- Crawler architecture
  - Google prototype [Page et al. 1996]
  - Mercator crawler [Heydon et al. 1999]
  - Polytech university [Shkapenyuk et al. 2002]
Summary

- Issues of parallel crawlers
  - Evaluation metrics
- Design alternatives
  - Crawler models
  - Experimental comparison
- Batching significantly reduces communication overhead and keeps high quality
- Many more details in the paper