Stanford WebBase Components and Applications

Junghoo Cho, Hector Garcia-Molina, Taher Haveliwala, Wang Lam, Andreas Paepcke, Sriram Raghavan, and Gary Wesley

9 September 2004
WebBase Goals

- Enable large-scale Web-related research
- Crawl and store significant pieces of the Web over time
- Index Web data easily and efficiently
- Clean, simple data retrieval over the network
Challenges

- **Crawling**
  - Minimize load on Web servers; honor robots.txt
  - Maximize throughput (parallelization)
  - Simplify operation (easy to start/stop/resume)

- **Indexing**
  - Scale to large datasets (storage, time)

- **Distribution**
  - Maximize throughput
  - Simplify retrieval (fetch and use of data)
Architecture
Site Crawler

- Crawl by site (FQDN)
- Independent units of work
- Coverage
- Database-backed controller
- Prevents excess crawl
- Site-specific crawl parameters: pages, depth, pause, types, ...
Crawler Performance

![Graph showing the relationship between the number of crawler processes and throughput (MB/s). The graph displays two lines: one for pages/sec and another for megabytes/sec. As the number of crawler processes increases, the throughput also increases.]
Crawler Performance

The graph shows the relationship between the number of crawler processes and the number of Web sites crawled concurrently. As the number of crawler processes increases, the number of Web sites crawled concurrently also increases linearly.
Distributor

Launch Distributors

Repository

DistribDaemon

WebBase server

stream request

User X

Client RunHandlers 1

Client RunHandlers 2

User Y

Client RunHandlers 1

External Users

Stanford
Request Interface

• Crawl
  – Choose a specific crawl space and date of interest

• Site or sequence of Web sites
  – Request specific sites, or all of them in crawl order

• Number of pages
  – Restrict fetch of site to a shallow crawl

• Opaque page offsets
  – Restart from a known point
Distributor Performance

scenario

throughput in pages/s

local-direct-compressed
local-direct-uncompressed
local-distributor-compressed
local-distributor-uncompressed
remote-compressed-1CPU
remote-compressed-2CPU
For More Information

- In submission: http://dbpubs.stanford.edu/pub/2004-34