INTERNET ARCHIVE

HERITRIX

Heritrix Crawler Update

IWAW 2005

Michael Stack

stack@ archive.org

Reporting on behalf of the Internet Archive Web Team and Heritrix community at large.

Internet Archive, The Presidio, San Francisco, CA

What is Heritrix?

- Goal: open-source, extensible, web-scale, archival-quality, web crawling software
- Collaboratively developed
  - Internet Archive, IIPC, partner libraries, and others
  - Over seven releases since January 2004
- Technical parameters:
  - Java, built on many other open source libraries
  - Tested and supported on Linux
  - Web-based control console
  - Highly configurable and customizable

Last year (IWAW'04)…

- Heritrix introduced at IWAW'04
  - Version 1.0
  - Core architecture, basic features in place
- Primarily useful for focused crawls
  - Hundreds to thousands of specific target websites
  - Over 20 million collected URIs per crawl
  - Crawls run for up to a week
  - Larger crawls hit single-machine resource limits (OOMEs).

Heritrix Releases: 1.2

- Heritrix 1.2: November 2004

  Motif: “requested features and fixes”

  – Better session-id handling (“URI canonicalization”)
  – IP politeness
  – More flexible scope (“SURT prefixes”)
  – Trial use of Berkeley DB (Java) for oversized data structures

Heritrix Releases: 1.4

- Heritrix 1.4: April 2005

  Motif: “memory efficiency at scale”

  – Much improved single-machine capacity
    - Big data structures moved to BDB JE.
  – Customizable 'decide rule' scope options
  – Improved balanced-progress and junk-control (“queue budgeting”)
  – Experimental “Adaptive Revisit” frontier (by Kristinn Sigurðsson, National Library of Iceland)
  – Experimental programmatic remote control (via “JMX” Java management extensions)

Heritrix: Larger Crawls

- .GOV for US National Archives (NARA)
  - 5 weeks through November 2004
  - Approximately 1,400 .GOV domains
  - 5 machines: 75 million URLs, 6.5 TB compressed
- .FR for Bibliotheque Nationale de France
  - 5 weeks through January 2005
  - Approximately 370k target .FR hosts
  - 13 machines; 100 million URLs, 3TB compressed
- .AU for National Library of Australia
  - 6 weeks through July 2005
  - Started with 340k seeds. Ended with 800k .AU & other hosts.
  - 3 machines; 190 million URLs, 6.7 TB uncompressed, 4.5 TB compressed

Heritrix: Current Work

- Heritrix 1.6: in progress; due October

  Motif: “manageability at scale”

  – Expanded JMX monitoring & control

  – Checkpoint & resume crawls

  – Sustained performance >30 million URLs crawled.
    - Bloom filter (Formerly BDB JE).

  – Experimental tools to help in multi-machine coordination
    - Crawl Splitter Processor
    - Cluster controller

INTERNET ARCHIVE

Heritrix: Future Plans (2.0+)

- Bigger & faster crawling
  - Automatic multi-machine coordination
  - Improved prioritization in limited time/space budgets
- Better junk, spam, & duplicate-content heuristics
  - Constant research and improvement needed
  - Automate expert operator judgments, enable bulk adjustments mid-crawl
- Continuous revisit crawling at scale
  - Minimize storage costs of unchanged content
- Support new protocols & formats
  - Streaming Media, etc.
  - Write WARC format
- Continued growth of community around Heritrix.
  - Currently, linguists, web scientists, and 'artists'.

INTERNET ARCHIVE

Das Ende

HERITRIX

Thank you

stack@ archive.org