Searching Web Archive Collections

Michael Stack
Internet Archive Web Team
The Presidio of San Francisco
4 Funston Ave.
San Francisco, CA 94129
stack@archive.org

Abstract

Web

A single URL may appear multiple times in a WAC. Each instance may differ radically, minimally or not at all across crawls, harvested by other means, the Nutch fetcher step had to be recast to instead pull content from the WAC repository rather than from the live web. At IA and other institutions using the Heritrix web crawler [6], harvested content is stored in the ARC file format [7]; composite log files each with many collected URLs. For the IA, an ARC-to-segment tool was written to feed ARC files to the Nutch parsers and segment content writers. (Adaptation for formats other than IA ARC

Clicking on the 'RSS' image in the above returns an RSS representation of the search results [11];

distributing the parsing, update, and indexing work across a cluster needs

http://labs.google.com/papers/mapreduce-osdi04.pdf
[15] Petabox *http://www.archive.org/web/petabox.php*