Refactoring the ht://Dig Search Engine

WWW 2007
May 12, 2007

Neal Richter
Anthony Arnone

RIGHT NOW TECHNOLOGIES

What is ht://Dig?

- Three Part Package
  - Robust crawler
  - Search index (BDB)
  - Search front end (CGI)
- Sourceforge Project
  - RightNow Sponsorship
  - LGPL
- Robust URL Handling
  - Many connection types supported
  - Index and search time filters
  - URL rewriting

RIGHT NOW TECHNOLOGIES

Motivation

- Spaghetti code
  - Hard to maintain
  - Inconsistent programming paradigms
  - Very hard to extend
- (Almost) No multilingual support
  - Accented text supported through addons
- Poor scalability
  - Inefficient inverted index schema
  - Frequent optimization steps needed

RIGHT NOW TECHNOLOGIES

Modernization

- Search index
  - CLucene
  - Dual index design
- HTML parser
  - HTMLTidy (tidylib)
- UTF-8 support
  - CLucene uses UCS2
- Modularity
  - Parsers and search index broken out into APIs
  - Well-defined code paths
- Modern language features
  - Standard template library

RIGHT NOW TECHNOLOGIES

Added Features

- Utility library
  - C/C++ indexing API
  - C/C++ and PHP searching API
- Lucene features
  - Boolean, phrase, prefix, etc… queries
  - Text analyzers
  - Easy to add fields
- Single document insertion
  - Can provide text to augment HTML
  - Document or URL
- Sitemap and extended sitemap
  - Extended sitemap textFacets

RIGHT NOW TECHNOLOGIES

Future Work

- Tools collection
  - Index monitoring and tweaking
  - Luke
- Command line interface for crawling
- Search front end
  - Provide XML or form based search
  - Public interface for common languages
- Further modularization
- Solr / Nutch integration

RIGHT NOW TECHNOLOGIES

Conclusion

- Provide updated features to a project with a solid (though arcane) codebase
- Allow easier future additions
- Modern text search engine
- Lots more work to be done

RIGHT NOW TECHNOLOGIES