Nutch2Roadmap

Nutch2Roadmap

Introduction

This page is designed to provide a list of the features and architectural changes that will be implemented in Nutch 2.X. It is important to recognize:

- this document is meant to serve as a basis for discussion, feel free to contribute to it
- many aspects of this document may also serve relevance and also feature on the 1.X codebase

Proposed Tasks

- Giraph support. There is an existing implementation for Nutch 2.X but it needs to be revisited
- Sitemap support using Crawler Commons
- HTML5 support
- RDF Microformats Support
- Static Snippet Generation
- Sentences Detection and Named Entity Recognize
- Plugin cleanup : Tika only for parsing document formats (see http://wiki.apache.org/nutch/TikaPlugin)
    - keep only stuff HtmlParseFilters (probably with a different API) so that we can post-process the DOM created in Tika from whatever original format.
    - Modify code so that parser can generate multiple documents which is what 1.x does but not 2.0
    - Offload url filtering and url normalization, URL state management, perhaps deduplication to [http://code.google.com/p/crawler-commons/]. We should coordinate our efforts, and share code freely so that other projects (bixo, heritrix,droids) may contribute to this shared pool of functionality, much like Tika does for the common need of parsing complex formats.
- Rewrite SOLR deduplication : do everything using the webtable and avoid retrieving content from Solr
- canonical tag support
- better handling of redirects
- detecting duplicated sites
- detection of spam cliques
- additional tools to manage the webgraph

Completed Tasks

- Hadoop 2.x support (This depends to Gora)
- Nutch 2.X Docker Container - Containers for various Nutch 2.X configurations
- ~~Storage Abstraction~~
    - ~~initially with back end implementations for HBase and HDFS~~
    - ~~extend it to other storages later e.g. MySQL etc...~~
- ~~-Externalize functionality to crawler-commons project [http://code.google.com/p/crawler-commons/] starting with robots handling~~
- -Remove index / search and delegate to SOLR -
    - ~~-we may still keep a thin abstract layer to allow other indexing/search backends (ElasticSearch?), but the current mess of indexing/query filters and competing indexing frameworks (lucene, fields, solr) should go away. We should go directly from DOM to a NutchDocument, and stop there.~~