NewScoringIndexingExample

Example Running new Scoring and Indexing Systems

N.B. This page and the functionality described within is only applicable and relevant to Nutch 1.X.

Introduction

Below is an example of running the new scoring and indexing systems from start to finish. This was done with a sample of 1000 urls and I ran two different fetch cycles. The first being 1000 urls and the second being the top 2000 urls. The loops job is optional but included for completeness. In production we have actually removed that job. This was done with a clean pull from Nutch trunk as of 2009-03-06 (right before 1.0 is set to be released). If anybody has any problems running these commands or has questions send me an email or send one to the nutch users or dev list and I will reply. Please send it to kubes at the apache address dot org.

Workflow

```
bin/nutch inject crawl/crawldb crawl/urls/
bin/nutch generate crawl/crawldb/ crawl/segments
bin/nutch fetch crawl/segments/20090306093949/
bin/nutch updatedb crawl/crawldb/ crawl/segments/20090306093949/
bin/nutch org.apache.nutch.scoring.webgraph.WebGraph -segment crawl/segments/20090306093949/ -webgraphdb crawl
/webgraphdb
```

One thing to point out here is that WebGraph is meant to be used on larger web crawls to create web graphs. By default it ignores outlinks to pages in the same domain, including subdomains, and pages with the same hostname. It also limits to one outlink per page to links in the same page or the same domain. All of these options are changeable through the following configuration options:

Configuration

```
<!-- linkrank scoring properties -->
<property>
  <name>link.ignore.internal.host</name>
  <value>true</value>
  <description>Ignore outlinks to the same hostname.</description>
</property>

<property>
  <name>link.ignore.internal.domain</name>
  <value>true</value>
  <description>Ignore outlinks to the same domain.</description>
</property>

<property>
  <name>link.ignore.limit.page</name>
  <value>true</value>
  <description>Limit to only a single outlink to the same page.</description>
</property>

<property>
  <name>link.ignore.limit.domain</name>
  <value>true</value>
  <description>Limit to only a single outlink to the same domain.</description>
</property>
```

Additional WebGraph Classes

But by default if you are only crawling pages within a domain or within a set of subdomains, all outlinks will be ignored and you will come up with an empty webgraph. This in turn will throw an error while processing through the LinkRank job. The flip side is by NOT ignoring links to the same domain/host and by not limiting those links, the webgraph becomes much, much more dense and hence there is a lot more links to process which probably won't affect relevancy as much.

```
bin/nutch org.apache.nutch.scoring.webgraph.Loops -webgraphdb crawl/webgraphdb/
bin/nutch org.apache.nutch.scoring.webgraph.LinkRank -webgraphdb crawl/webgraphdb/
bin/nutch org.apache.nutch.scoring.webgraph.ScoreUpdater -crawldb crawl/crawldb -webgraphdb crawl/webgraphdb/
bin/nutch org.apache.nutch.scoring.webgraph.NodeDumper -scores -topn 1000 -webgraphdb crawl/webgraphdb/ -output
crawl/webgraphdb/dump/scores


more crawl/webgraphdb/dump/scores/part-00000
--------------------------------------------------------------------------------
http://validator.w3.org/check?uri=referer        0.4955311
http://www.adobe.com/go/getflashplayer        0.4060498
http://www.statcounter.com/        0.4060498
http://www.liveinternet.ru/click        0.33680826
http://www.adobe.com/products/acrobat/readstep2.html        0.31656843
http://www.adobe.com/go/getflashplayer/        0.30378538
http://www.bloomingbows.com/2003/scripts/sitemap.asp        0.27821928
http://www.misterping.com/        0.27821928
...
--------------------------------------------------------------------------------


bin/nutch readdb crawl/crawldb/ -stats
--------------------------------------------------------------------------------
CrawlDb statistics start: crawl/crawldb/
Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
Statistics for CrawlDb: crawl/crawldb/
TOTAL urls:        16711
retry 0:        16686
retry 1:        25
min score:        0.0
avg score:        0.022716654
max score:        0.495
status 1 (db_unfetched):        15739
status 2 (db_fetched):        677
status 3 (db_gone):        75
status 4 (db_redir_temp):        143
status 5 (db_redir_perm):        77
CrawlDb statistics: done
--------------------------------------------------------------------------------


bin/nutch generate crawl/crawldb/ crawl/segments/ -topN 2000
bin/nutch fetch crawl/segments/20090306100055/
bin/nutch updatedb crawl/crawldb/ crawl/segments/20090306100055/
rm -fr crawl/webgraphdb/
bin/nutch org.apache.nutch.scoring.webgraph.WebGraph -segment crawl/segments/20090306093949/ -segment crawl
/segments/20090306100055/ -webgraphdb crawl/webgraphdb
```

One thing that has been brought up is the -segment flag on webgraph. If you have more than one segment then you would use the -segmentDir flag available on the command line interface.

```
bin/nutch org.apache.nutch.scoring.webgraph.Loops -webgraphdb crawl/webgraphdb/
bin/nutch org.apache.nutch.scoring.webgraph.LinkRank -webgraphdb crawl/webgraphdb/
bin/nutch org.apache.nutch.scoring.webgraph.ScoreUpdater -crawldb crawl/crawldb -webgraphdb crawl/webgraphdb/

more crawl/webgraphdb/dump/scores/part-00000
--------------------------------------------------------------------------------
http://www.statcounter.com/        1.7133079
http://www.morristownwebdesign.com/        1.0093393
http://www.jdoqocy.com/click-3331968-10419685        0.87828785
http://www.anrdoezrs.net/click-3331968-10384568        0.87828785
http://www.sedo.com/main.php3?language=e        0.6565905
http://wetter.spiegel.de/spiegel/html/frankreich0.html        0.641775
http://www.kenwood.com/        0.6084726
http://validator.w3.org/check?uri=referer        0.5605916
http://wetter.spiegel.de/spiegel/html/Italien0.html        0.5164927
http://www.youtube.com/?hl=en&tab=w1        0.50952965
http://www.addthis.com/bookmark.php        0.5013165
http://www.ptguide.com/        0.49564213
http://www.adobe.com/go/getflashplayer        0.47368217
http://de.weather.yahoo.com/ITXX/ITXX0073/index_c.html        0.4657473
http://www.adobe.com/shockwave/download/download.cgi?P1_Prod_Version=ShockwaveFlash&promoid=BIOW
0.44376293
http://www.google.com/        0.42282072
http://www.zajezdy.cz/        0.41620353
http://www.intermarche.com/        0.41489196
http://www.shipskill.com/7/        0.4147887
http://www.statcounter.com/free_hit_counter.html        0.40928197
http://www.skandinavien-fans.de/frameset.html        0.40886167
http://ff.connextra.com/Bet365/selector/click?client=Bet365&placement=Livescore_NS4default_133x200
0.39578557
http://deluxe-menu.com/        0.3891917
http://www.entreprises.gouv.fr/zerocharges/        0.38621464
http://www.macromedia.com/go/getflashplayer        0.38621464
http://www.nhlbi.nih.gov/whi/        0.38621464
http://www.microsoft.com/windows/ie/        0.38621464
http://www.quantcast.com/p-65DrxcUXjcWq6        0.38621464
http://janvanderperk.write2me.nl/        0.38621464
...
--------------------------------------------------------------------------------


bin/nutch readdb crawl/crawldb/ -stats
--------------------------------------------------------------------------------
TOTAL urls:        55730
retry 0:        55641
retry 1:        88
retry 2:        1
min score:        0.0
avg score:        0.020518823
max score:        1.713
status 1 (db_unfetched):        52824
status 2 (db_fetched):        2091
status 3 (db_gone):        229
status 4 (db_redir_temp):        374
status 5 (db_redir_perm):        212
CrawlDb statistics: done
--------------------------------------------------------------------------------


bin/nutch org.apache.nutch.indexer.field.BasicFields -output crawl/fields/basicfields -segment crawl/segments
/20090306093949/ -segment crawl/segments/20090306100055/ -webgraphdb crawl/webgraphdb/
bin/nutch org.apache.nutch.indexer.field.AnchorFields -basicfields crawl/fields/basicfields/ -output crawl
/fields/anchorfields -webgraphdb crawl/webgraphdb/
bin/nutch org.apache.nutch.indexer.field.FieldIndexer -fields crawl/fields/basicfields/ -fields crawl/fields
/anchorfields/ -output crawl/indexes
```

Class Diagram

Below is a thumbnail Class Diagram representing the Java Class ecosystem for WebGraph. You can click on the thumbnail for a much larger, downloadable picture.

NutchWebGraph.png]]