MetacatUI - Bug #6044

Enable sub-text searching in Solr queries

08/01/2013 04:45 PM - Chris Jones

Status:	In Progress	**Start date:**	08/01/2013
Priority:	Normal	**Due date:**	
Assignee:	Chris Jones	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Bugzilla-Id:			

Description

When submitting a search to the Solr index, we are only getting hits for whole-word searches. We want to be able to return results for fragments of text, like, a search for 'ocean' would also return hits of 'oceanographic'.

History

#1 - 08/01/2013 05:23 PM - Chris Jones

In researching this, it looks like we'll need to change the Solr schema to use a different type of analyzer than the standard whitespace-delimited analyzer. We can chain the analyzers together, and use the NGramFilterFactory, or potentially the EdgeNGramFilterFactory. These filters will analyze text fields in the documents, and will decompose them to partial words at a length specified by the minGramSize and maxGramSize parameters. So, for the word 'oceanographic', 'ocean' would be a hit for EdgeNGramFilterFactory with a min gram size of 5, 4, 3, etc. To match the sub-term of 'graphic', we'd need the NGramFilterFactory, with a min gram size of no more than 7.

The impact of these filters on the index is that it will increase the number of indexed words many fold. I would think an NGram min size of 3 would be the shortest word we would want to match, possibly 4 or 5.

Changing the schema filters will require a re-index of the database contents.

#2 - 08/02/2013 01:13 PM - Chris Jones

- Target version changed from 1.0.0 to 1.1.0

Moving this to 1.1.0. The filters above might really bloat the index, and the current filters in schema.xml allow for * searches as a wildcard, but we're finding inconsistent results. *oil seems to match oil and soil, but *henology doesn't match phenology. Trailing *s seem to work better. This needs more investigation.

#3 - 08/23/2013 03:55 PM - ben leinfelder

- Target version changed from 1.1.0 to 1.2.0

#4 - 10/18/2013 04:10 PM - ben leinfelder

- Target version changed from 1.2.0 to 1.3.0

#5 - 11/14/2013 08:09 AM - ben leinfelder

- Target version deleted (1.3.0)