12 Underutilized Features of Apache Lucene and Solr

lucid IMAGINATION

Thinking Lucene ◥ Think Lucid.

It's that time of year

- **In the spirit of the 12 Days of Christmas, eh!**
- **Don't worry, I won't sing**
- **These are not in any particular order and are only my opinion!**
  - In other words, it's not like I did a scientific poll (but I did poll some other committers)

Queries

- **There are over 40 classes in Lucene/Solr that extend the Query class**
  - Phrases (automatically for any query with multiple words)
  - Spans
  - Function/CustomScore
  - Many more. Take time to learn what they do.

- **Bonus**
  - New Automaton queries

Field Selector

- **Retrieve only the fields you need from storage**
  - IndexReader.document(int, FieldSelector)
- **Use Case: 1-2 large fields, bunch of small fields**
- **Several options, including lazy loading**

Deeper Analysis/Customization

- **Custom Tokenizers/TokenFilters**
  - Lucene user's especially should consider writing their own Analyzer
- **Payloads**
- **Named Entity Extraction**
  - OpenNLP, others
- **Language handling, especially for Asian languages**

Non Default Settings

- **Solr users are especially guilty of simply copying schema.xml and solrconfig.xml with nary an eye to whether it is right for them**

- **Defaults you may or may not like:**
  - OR Operator, Merge Factor, RAM Buffer Size in MB, Compound File, “StandardAnalyzer”, Lucene Query Parser, Max Field Length
- **Other**
  - Spellchecking, more like this, highlighting

Logs, Users and Relevance Tuning

- **IndexWriter.setInfoStream() and SLF4J (Solr)**
- **When was the last time you looked at your top 50 user queries?**
  - How many queries return zero results and why?
- **IndexReader.explain() and &debugQuery=true**

Alternate Query Parsers

- Surround
- XML
- Complex Phrase
- Dismax/eDismax (Solr)
- QParserPlugins: Boost, Spatial (next release), Raw, others
  - 15 total in trunk

Query Expansion

- Synonyms
- More Like This
- MultiTermQuery: fuzzy, wildcard, etc.
- Caveats

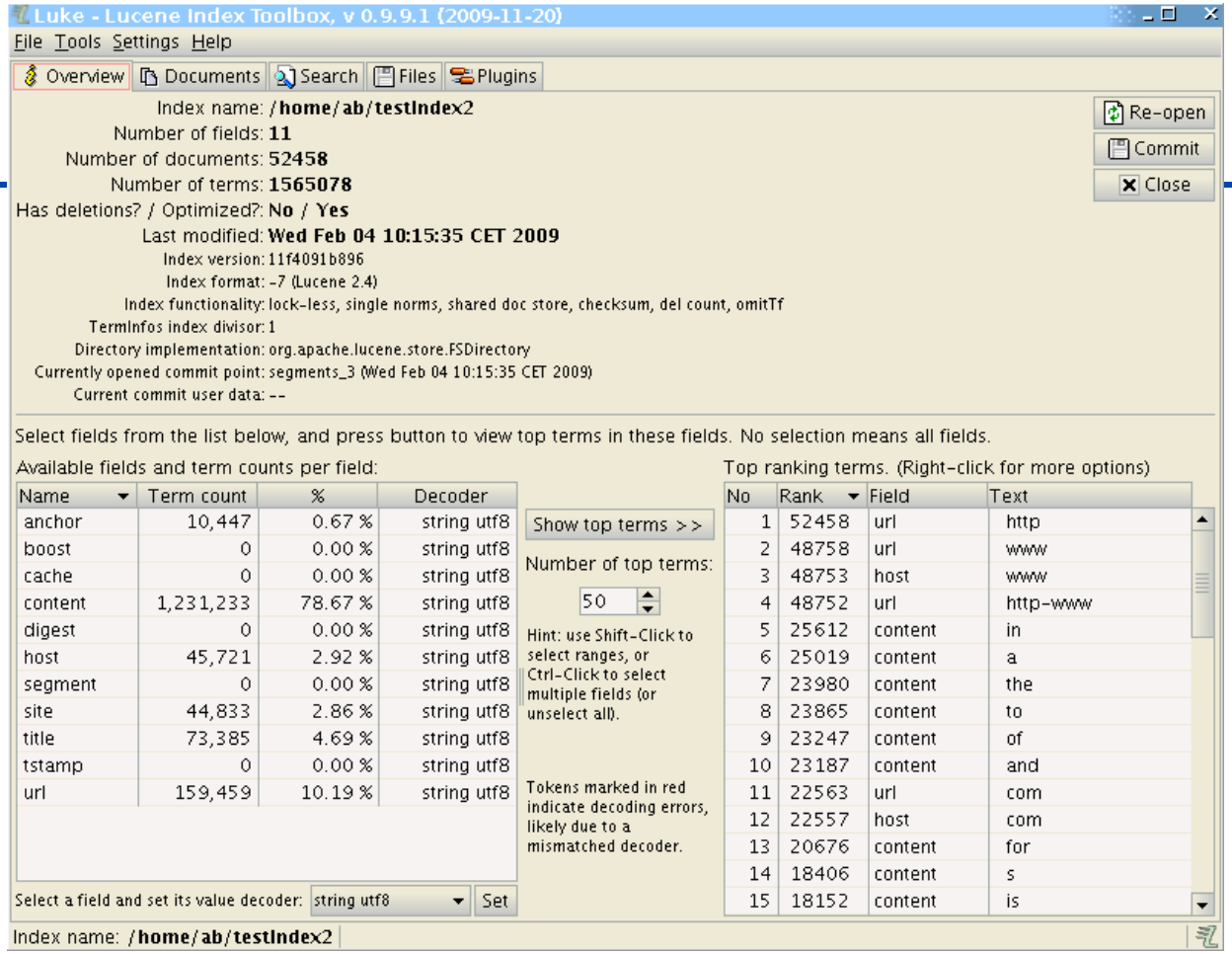
Near Real Time Indexing/Searching

- IndexReader.reopen()
- IndexReader.open(IndexWriter)
- Warming of IndexWriter segments

JIRA and the Mail Archives

- Chances are your problem has been discussed before
  - http://search.lucidimagination.com
  - https://issues.apache.org/jira/browse/LUCENE
  - https://issues.apache.org/jira/browse/SOLR

Luke



- http://code.google.com/p/luke/
- http://wiki.apache.org/solr/LukeRequestHandler
- Solr's Schema Browser

Fake and Invisible Queries

- **Caching**
- **Canned Results**
  - QueryElementationComponent
- **Multiple internal queries per user query**
  - Fallback queries

Trunk

- **Without trunk users, trunk releases take longer**
- **A word on versions:**
  - Released
  - Branch_3x
  - Trunk
- **We strive to keep trunk/Branch_3x stable**
- **Without users trying and reporting on experiences on trunk, progress slows**
- **There are some seriously cool features on trunk right now**

Bonus: Plugins and Extensions

- Contrib/Benchmark
- SearchComponents
- UpdateRequestProcessor
- Similarity
  - SweetSpotSimilarity
- Clustering (Carrot^2)