



12 Underutilized Features of Apache Lucene and Solr

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

Index name: /home/ab/testIndex2
Number of fields: 11
Number of documents: 52458
Number of terms: 1565078
Has deletions? / Optimized?: No / Yes
Last modified: Wed Feb 04 10:15:35 CET 2009
Index version: 11f4091b896
Index format: -7 (Lucene 2.4)
Index functionality: lock-less, single norms, shared doc store, checksum, del count, omitTF
Terminfos index divisor: 1
Directory implementation: org.apache.lucene.store.FSDirectory
Currently opened commit point: segments_3 (Wed Feb 04 10:15:35 CET 2009)
Current commit user data: --

Select fields from the list below, and press button to view top terms in these fields. No selection means all fields.

Available fields and term counts per field:

Name	Term count	%	Decoder
anchor	10,447	0.67 %	string utf8
boost	0	0.00 %	string utf8
cache	0	0.00 %	string utf8
content	1,231,233	78.67 %	string utf8
digest	0	0.00 %	string utf8
host	45,721	2.92 %	string utf8
segment	0	0.00 %	string utf8
site	44,833	2.86 %	string utf8
title	73,385	4.69 %	string utf8
tstamp	0	0.00 %	string utf8
url	159,459	10.19 %	string utf8

Show top terms >>

Number of top terms: 50

Hint: use Shift-Click to select ranges, or Ctrl-Click to select multiple fields (or unselect all).
Tokens marked in red indicate decoding errors, likely due to a mismatched decoder.

Select a field and set its value decoder: string utf8 Set

Top ranking terms. (Right-click for more options)

No	Rank	Field	Text
1	52458	url	http
2	48758	url	www
3	48753	host	www
4	48752	url	http-www
5	25612	content	in
6	25019	content	a
7	23980	content	the
8	23865	content	to
9	23247	content	of
10	23187	content	and
11	22563	url	com
12	22557	host	com
13	20676	content	for
14	18406	content	s
15	18152	content	is

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

Fake and Invisible Queries

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

- **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 (Carrot²)**