Lucene's Latest (for Libraries)

Erik Hatcher, LucidWorks
Or...

"what I learned last night"
Agenda

- Lucene powers the search capabilities of practically all library discovery platforms, by way of Solr, etc. The Lucene project evolves rapidly, and it's a full-time job to keep up with the ever improving features and scalability. This talk will distill and showcase the most relevant(!) advancements to date.
About me

- "Old Timer", apparently
- Yet, barely knows MARC from Adam
Where we left off

- [http://code4lib.org/conference/2013/hatcher](http://code4lib.org/conference/2013/hatcher)

Solr is continually improving. Solr 4 was recently released, bringing dramatic changes in the underlying Lucene library and Solr-level features. It's tough for us all to keep up with the various versions and capabilities.

This talk will blaze through the highlights of new features and improvements in Solr 4 (and up). Topics will include: SolrCloud, direct spell checking, surround query parser, and many other features. We will focus on the features library coders really need to know about.
release_version:[4.1 TO *]

- 4.7: 26 February 2014, 4.7.1 any day now
- 4.6: 24 November, 4.6.1: 28 January 2014
- 4.5: 5 October, 4.5.1: 24 October
- 29 July 2013 - Apache Solr Reference Guide 4.4 Available
- 4.4: 23 July
- 4.3: 6 May, 4.3.1: 18 June
- 4.2: 13 March, 4.2.1: 3 April
- #c4l13: 13 February
- 4.1: 22 January 2013
Suggester

- SOLR-5378: "Suggester Version 2"
- Dictionary pluggability
- Map the suggester options
- "beefier" Lookup support instead of resorting to collation and such. (Move computation from query time to index time) with more freedom
- AnalyzingInfixLookupFactory, BlendedInfixSuggester
- FreeTextSuggester
- Payload and expression support
Suggester

- AnalyzingInfixLookupFactory (4.6)

- Suggest improvements: a new SuggestComponent that fully utilizes the Lucene suggester module; queries can now use multiple suggesters; Lucene's FreeTextSuggester and BlendedInfixSuggester are now supported. (4.7)

- Analyzing/FuzzySuggester now allow to record arbitrary byte[] as a payload (4.3)

- FreeTextSuggester: can predict the next word using a simple ngram language model useful for "long tail" suggestions. (4.6)

- BlendedInfixSuggester: like AnalyzingInfixSuggester but boosts suggestions that matched tokens with lower positions. (4.7)

- http://blog.mikemccandless.com/2013/06/a-new-lucene-suggester-based-on-infix.html
Expressions

- SOLR-5378 (4.7):
  - `<str name="weightExpression">((price * 2) + ln(popularity))</str>`

- SOLR-5707: Lucene Expressions in Solr
  - In-progress
Configuration API!

- REST (4.2+...)
  - Schema: fields, copyFields

- Schemaless mode: Added support for a mode that requires no up-front schema modifications, in which previously unknown fields' types are guessed based on the values in added/updated documents, and are then added to the schema prior to processing the update. Note that the below-described features are also useful independently from schemaless mode operation. (4.4)

- AddSchemaFieldsUpdateProcessor,
  Parse*UpdateProcessorFactory

- Across all configuration, SOLR-5653: work in progress
SolrCloud

- Collection API improvements
- Shard splitting (4.3)
- Custom sharding support, including the ability to shard by field. (4.5)
- CloudSolrServer
  - can now route updates directly to the appropriate shard leader. (4.5)
- SSL (4.7)
Querying

- New MaxScoreQParserPlugin: Return max() instead of sum() of terms: {!maxscore} (4.4)
- New CollapsingQParserPlugin for high performance field collapsing on high cardinality fields (4.6)
- Add a Lucene and Solr QParserPlugin for Lucene's SimpleQueryParser: {!simple} (4.7)
- Significant performance improvements for minShouldMatch (mm) queries due to skipping resulting in up to 4000% faster queries. (4.3)
Miscellaneous

- Added a new classification module (4.2)
- PostingsHighlighter now allows custom passage scores, and other improvements (4.3)
- Analyzing/FuzzySuggester now allow to record arbitrary byte[] as a payload. The suggesters also use an ending offset to determine whether the last token was finished or not, so that a query "i " will no longer suggest "Isla de Muerta" for example. (4.3)
- PatternCaptureGroupTokenFilter (4.4)
http://searchhub.org/2013/06/27/poor-mans-entity-extraction-with-solr/

```xml
<field name="acronyms" type="caps" indexed="true" stored="false" multiValued="true"/>
<copyField source="content" dest="acronyms"/>

<fieldType name="caps" class="solr.TextField" sortMissingLast="true">
    <analyzer>
        <tokenizer class="solr.KeywordTokenizerFactory"/>
        <filter class="solr.PatternCaptureGroupFilterFactory" pattern="((?:[A-Z]\.?){3,})" preserve_original="false"/>
    </analyzer>
</fieldType>
```
Miscellaneous

- **Spatial**
  - Spatial queries can now search for indexed shapes by "IsWithin", "Contains" and "IsDisjointTo" relationships, in addition to typical "Intersects". (4.3)
  - Upgrade to Spatial4j 0.4. Various new options are now exposed automatically for an RPT field type (4.7)

- **Faceting**
  - Multithreaded faceting. (4.5)
  - Faceting now supports local parameters for faceting on the same field with different options. (4.3)
## Miscellaneous

- Core discovery, rather than mandatory solr.xml (4.3)
  - core.properties (4.4)
- Various new highlighting configuration parameters. (4.3)
- Added a new system wide info admin handler that exposes the system info that could previously only be retrieved using a SolrCore. (4.4)
- The CSV Update Handler now supports optionally adding the line number/row id to a document. (4.4)
- EnumField (4.6)
- Solr indexes and transaction logs may stored in HDFS with full read/write capability. (4.4)
Scale

- New cursorMark request param for efficient deep paging of sorted result sets. See http://s.apache.org/cursorpagination (4.7)
- Add a Solr contrib that allows for building Solr indexes via Hadoop's MapReduce (4.7)
- DocValues in Solr (4.2)
  - DocValue improvements: single valued fields no longer require a default value, allowing dynamicFields to contain doc values, as well as sortMissingFirst and sortMissingLast on docValue fields. (4.5)
Block join

- SOLR-3076 (4.4): Block joins. Documents and their sub-documents must be indexed as a block.
  - `{!parent which=<allParents>} <someChildren>` takes in a query that matches child documents and results in matches on their parents.
  - `{!child of=<allParents>} <someParents>` takes in a query that matches some parent documents and results in matches on their children.
Java 7

- IMPORTANT IF YOU'RE RUNNING JAVA 6!!!
- Apache Lucene/Solr 4.8 will require Java 7
  - [announced 12 March 2014]
Keep an eye on...

- SOLR-5302: Analytics Component (in progress on trunk)
- LUCENE-5207, LUCENE-5334: Added expressions module for customizing ranking with script-like syntax (4.6)
  - In-progress: SOLR-5707
- SOLR-5720: Add ExpandComponent to expand results collapsed by the CollapsingQParserPlugin (4.8, committed to 4x after 4.7.1)
- LUCENE-4518: Suggesters: highlighting (explicit markup of user-typed portions vs. generated portions in a suggestion) [EBSCO?]
- SOLR-4787: Join contrib: hjoin, bjoin, and vjoin (4.8, maybe?)
*bows*

- ARP, NINES, Rossetti Archive, Collex
  - Bethany Nowviskie, Jerome McGann
- Bess
- #code4lib
- Project Blacklight
- Lucene community: Mike McCandless, Robert Muir, Hoss, Erickson, Uwe...
- Lucid( Imagination | Works)
https://twitter.com/lisafreyer/status/448487103366971392
For more information...

- Mike McCandless
  - [http://blog.mikemccandless.com](http://blog.mikemccandless.com)
  - [http://jirasearch.mikemccandless.com](http://jirasearch.mikemccandless.com)

- Apache Lucene
  - CHANGES.txt, JIRA
  - Apache Solr Reference Guide (4.4+)

- LucidWorks
  - [http://www.lucidworks.com](http://www.lucidworks.com)