

Solr 1.4 Enterprise Search Server

Enhance your search with faceted navigation, result highlighting, fuzzy queries, ranked scoring, and more

David Smiley

Eric Pugh



BIRMINGHAM - MUMBAI

ULB Darmstadt



17347888

Table of Contents

Preface	1
Chapter 1: Quick Starting Solr	7
An introduction to Solr	7
Lucene, the underlying engine	8
Solr, the Server-ization of Lucene	8
Comparison to database technology	9
Getting started	10
The last official release or fresh code from source control	11
Testing and building Solr	12
Solr's installation directory structure	13
Solr's home directory	15
How Solr finds its home	15
Deploying and running Solr	17
A quick tour of Solr!	18
Loading sample data	20
A simple query	22
Some statistics	24
The schema and configuration files	25
Solr resources outside this book	26
Summary	27
Chapter 2: Schema and Text Analysis	29
MusicBrainz.org	30
One combined index or multiple indices	31
Problems with using a single combined index	33
Schema design	34
Step 1: Determine which searches are going to be powered by Solr	35
Step 2: Determine the entities returned from each search	35

Step 3: Denormalize related data	36
Denormalizing—"one-to-one" associated data	36
Denormalizing—"one-to-many" associated data	36
Step 4: (Optional) Omit the inclusion of fields only used in search results	38
The schema.xml file	39
Field types	40
Field options	40
Field definitions	42
Sorting	44
Dynamic fields	45
Using copyField	46
Remaining schema.xml settings	47
Text analysis	47
Configuration	48
Experimenting with text analysis	50
Tokenization	52
WorkDelimiterFilterFactory	53
Stemming	54
Synonyms	55
Index-time versus Query-time, and to expand or not	57
Stop words	57
Phonetic sounds-like analysis	58
Partial/Substring indexing	60
N-gramming costs	61
Miscellaneous analyzers	62
Summary	63
Chapter 3: Indexing Data	65
Communicating with Solr	65
Direct HTTP or a convenient client API	65
Data streamed remotely or from Solr's filesystem	66
Data formats	66
Using curl to interact with Solr	66
Remote streaming	68
Sending XML to Solr	69
Deleting documents	70
Commit, optimize, and rollback	70
Sending CSV to Solr	72
Configuration options	73
Direct database and XML import	74
Getting started with DIH	75
The DIH development console	76

DIH documents, entities	78
DIH fields and transformers	79
Importing with DIH	80
Indexing documents with Solr Cell	81
Extracting binary content	81
Configuring Solr	83
Extracting karaoke lyrics	83
Indexing richer documents	85
Summary	88
Chapter 4: Basic Searching	89
Your first search, a walk-through	89
Solr's generic XML structured data representation	92
Solr's XML response format	93
Parsing the URL	94
Query parameters	95
Parameters affecting the query	95
Result paging	96
Output related parameters	96
Diagnostic query parameters	98
Query syntax	99
Matching all the documents	99
Mandatory, prohibited, and optional clauses	99
Boolean operators	100
Sub-expressions (aka sub-queries)	101
Limitations of prohibited clauses in sub-expressions	102
Field qualifier	102
Phrase queries and term proximity	103
Wildcard queries	103
Fuzzy queries	105
Range queries	105
Date math	106
Score boosting	107
Existence (and non-existence) queries	107
Escaping special characters	108
Filtering	108
Sorting	109
Request handlers	110
Scoring	112
Query-time and index-time boosting	113
Troubleshooting scoring	113
Summary	115

Chapter 5: Enhanced Searching	117
Function queries	117
An example: Scores influenced by a lookupcount	118
Field references	120
Function reference	120
Mathematical primitives	121
Miscellaneous math	121
ord and rord	122
An example with scale() and lookupcount	123
Using logarithms	123
Using inverse reciprocals	124
Using reciprocals and rord with dates	126
Function query tips	128
Dismax Solr request handler	128
Lucene's DisjunctionMaxQuery	130
Configuring queried fields and boosts	131
Limited query syntax	131
Boosting: Automatic phrase boosting	132
Configuring automatic phrase boosting	133
Phrase slop configuration	134
Boosting: Boost queries	134
Boosting: Boost functions	137
Min-should-match	138
Basic rules	139
Multiple rules	139
What to choose	140
A default search	140
Faceting	141
A quick example: Faceting release types	142
MusicBrainz schema changes	144
Field requirements	146
Types of faceting	146
Faceting text	147
Alphabetic range bucketing (A-C, D-F, and so on)	148
Faceting dates	149
Date facet parameters	151
Faceting on arbitrary queries	152
Excluding filters	153
The solution: Local Params	155
Facet prefixing (term suggest)	156
Summary	158

Chapter 6: Search Components	159
About components	159
The highlighting component	161
A highlighting example	161
Highlighting configuration	163
Query elevation	166
Configuration	167
Spell checking	169
Schema configuration	169
Configuration in solrconfig.xml	171
Configuring spellcheckers (dictionaries)	173
Processing of the q parameter	175
Processing of the spellcheck.q parameter	176
Building the dictionary from its source	176
Issuing spellcheck requests	177
Example usage for a misspelled query	178
An alternative approach	180
The more-like-this search component	182
Configuration parameters	183
Parameters specific to the MLT search component	183
Parameters specific to the MLT request handler	184
Common MLT parameters	185
MLT results example	186
Stats component	189
Configuring the stats component	189
Statistics on track durations	190
Field collapsing	191
Configuring field collapsing	192
Other components	193
Terms component	194
termVector component	194
LocalSolr component	194
Summary	195
Chapter 7: Deployment	197
Implementation methodology	197
Questions to ask	198
Installing into a Servlet container	199
Differences between Servlet containers	199
Defining solr.home property	199

Logging	201
HTTP server request access logs	201
Solr application logging	203
Configuring logging output	203
Logging to Log4j	204
Jetty startup integration	205
Managing log levels at runtime	205
A SearchHandler per search interface	207
Solr cores	208
Configuring solr.xml	208
Managing cores	209
Why use multicore	210
JMX	212
Starting Solr with JMX	212
Take a walk on the wild side! Use JRuby to extract JMX information	215
Securing Solr	217
Limiting server access	217
Controlling JMX access	220
Securing index data	220
Controlling document access	221
Other things to look at	221
Summary	222
Chapter 8: Integrating Solr	223
Structure of included examples	223
Inventory of examples	224
SolrJ: Simple Java interface	224
Using Heritrix to download artist pages	226
Indexing HTML in Solr	227
SolrJ client API	230
Indexing POJOs	234
When should I use Embedded Solr	235
In-Process streaming	236
Rich clients	237
Upgrading from legacy Lucene	237
Using JavaScript to integrate Solr	238
Wait, what about security?	239
Building a Solr powered artists autocomplete widget with jQuery and JSONP	240
SolrJS: JavaScript interface to Solr	245
Accessing Solr from PHP applications	247
solr-php-client	248
Drupal options	250
Apache Solr Search integration module	251

Hosted Solr by Acquia	252
Ruby on Rails integrations	253
acts_as_solr	254
Setting up MyFaves project	255
Populating MyFaves relational database from Solr	256
Build Solr indexes from relational database	258
Complete MyFaves web site	260
Blacklight OPAC	263
Indexing MusicBrainz data	263
Customizing display	267
solr-ruby versus rsolr	269
Summary	270
Chapter 9: Scaling Solr	271
Tuning complex systems	271
Using Amazon EC2 to practice tuning	273
Firing up Solr on Amazon EC2	274
Optimizing a single Solr server (Scale High)	276
JVM configuration	277
HTTP caching	277
Solr caching	280
Tuning caches	281
Schema design considerations	282
Indexing strategies	283
Disable unique document checking	285
Commit/optimize factors	285
Enhancing faceting performance	286
Using term vectors	286
Improving phrase search performance	287
The solution: Shingling	287
Moving to multiple Solr servers (Scale Wide)	289
Script versus Java replication	289
Starting multiple Solr servers	290
Configuring replication	291
Distributing searches across slaves	291
Indexing into the master server	292
Configuring slaves	292
Distributing search queries across slaves	293
Sharding indexes	295
Assigning documents to shards	296
Searching across shards	297
Combining replication and sharding (Scale Deep)	298
Summary	300
Index	301
