

## はじめに

### 楽しい検索の世界へようこそ！

今日、検索を利用しない生活は想像ができなくらいに、身の回りに検索技術がありふれています。普段、私たちが何気なく検索しているところで、実は本書で紹介する Apache Solr (アパッチソーラー：以下 Solr と記します) の技術が利用されているかもしれません。

本書はそんな Solr を体系的に解説した書籍となります。Solr は Apache Lucene (アパッチルシーン：以下 Lucene と記します) を使って構築されたオープンソースの検索エンジンサーバです。Lucene は非常に優れたライブラリですが、利用するには Java のプログラムをたくさん書かなければいけません。Solr はその手間を解消するために HTTP でアクセスできる REST ライクなインタフェースを用意しています。また、インタフェースだけでなく、機能や使い勝手を向上するしくみがプラスされています。

前回の『Apache Solr 入門』の出版から約 3 年が経ちました。Lucene/Solr もバージョンが 4.5.1 (2013 年 10 月現在) となり、前のバージョン (Solr 1.4) に比べると大量データを扱うためのしくみや位置情報を活用した検索など、機能が大きく追加変更されています。これは、同様に Lucene をベースとした検索エンジン ElasticSearch が好敵手として登場しコミュニティが良い形で活性化されたこと、また、利用の拡大によりさまざまなニーズに対応してきたことの結果でしょう。

また、最近では Solr を核としたプロダクトや OSS (Cloudera Search や Yokozuna) も登場するなど、Solr 理解の重要性がより高まっている状況といえます。このように活発なコミュニティに支えられ、さまざまなシステムを支えるミドルウェアとなりつつある Solr に関して、このたび、ユーザの中でも特に経験豊かな執筆陣により、改訂版を出版できる運びとなりました。

本書が皆さまの理解に対する一助となることを願っています。

では、また本文でお会いしましょう！

2013 年 10 月  
大谷 純

### ●免責

本書に記載された内容は、情報の提供だけを目的としています。したがって、本書を用いた運用は、必ずお客様自身の責任と判断によって行ってください。これらの情報の運用の結果について、技術評論社および著者はいかなる責任も負いません。

本書記載の情報は、2013 年 10 月現在のものを掲載していますので、ご利用時には、変更されている場合もあります。

また、ソフトウェアに関する記述は、特に断りのないかぎり、2013 年 10 月現在のバージョンをもとにしています。ソフトウェアはバージョンアップされる場合があります。本書での説明とは機能内容や画面図などが異なってしまうこともあり得ます。本書ご購入の前に、必ずバージョン番号をご確認ください。

以上の注意事項をご承諾いただいた上で、本書をご利用願います。これらの注意事項をお読みいただかずにお問い合わせいただいても、技術評論社および著者は対処しかねます。あらかじめ、ご承知おきください。

### ●商標、登録商標について

・本書に登場する製品名などは、一般に各社の登録商標または商標です。なお、本文中に ™、® などのマークは特に記載していません。

## 謝辞

まず Solr の産みの親である Yonik Seeley さんに感謝いたします。Yonik さんはその類い希なるプログラミングの才を存分に発揮して、高性能検索エンジン Lucene を、Web 時代にマッチした軽快なインタフェースを通じて誰でも簡単に利用できるように仕立て上げました。そして CNET 社にも感謝いたします。Solr は CNET 社で開発されましたが、CNET 社が Apache にソースコードを寄贈してコードが公開されたことでユーザが一気に広がり、今日の Solr コミュニティの基礎を築きました。さらに、Christian Moen さんにも感謝いたします。Christian さんにより Kuromoji という日本語形態素解析ライブラリが寄贈されたことにより Solr でより簡単に日本語での検索が行えるようになりました。また、大谷個人としては、担当編集者の池本さんとリクルートテクノロジーの中野さん、ロンウィットの関口さん、Lucene/Solr コミュニティに感謝いたします。特に池本さん、中野さんは本書の改訂版を出版できるように尽力してくれました。関口さんには前回の『Apache Solr 入門』を書くきっかけと、いろいろと勉強させていただく機会を与えてもらっています。そして、Lucene/Solr コミュニティは、私に日々開発者 ML からのエキサイティングな英語のメールをプレゼントしてくれ、英語の勉強をする機会を与えてくれました。ありがとう！

## 対象読者

本書は、情報検索に興味を持つ、あらゆる人を対象としています。学生から社会人、週末プログラマから職業プログラマ、SE、プログラミングをしない情報システムを使うだけの人、検索エンジンを比較検討しようとしている人……制限はありません。それぞれの立場で Solr をお楽しみください。しかしながら、本書は Solr の技術解説書でもあり、Solr をインストールしたりサンプルコードを設定したりする場面もあります。そのときは技術的なバックグラウンドのある方は若干有利でしょう。自分の経験に感謝しつつ、コマンドを入力したり、検索して結果を確認したり、より深くお楽しみください。

## 本書の構成と読み方

本書は全 9 章から構成されますがすべてに目を通す必要はありません。Chapter 1 の前半では検索エンジンの基本を紹介しています。すでに検索エンジンについての知識があれば読まなくてもかまいません。次に Solr のインストール方法を説明しています。本書は Solr を使いながら説明しているところが多いので、ここでぜひ手元の PC に Solr を準備しておくことをお勧めします。最後にアーキテクチャを解説していますので、ここもぜひ目を通しておくといいでしょう。

Chapter 2 から Chapter 4 は Solr の基本知識である「スキーマ定義」「インデックス作成」および「検索」について体系的に説明しています。ぜひ一とおり読んでおくことをお勧めします。

Chapter 5 はプログラマのための章です。Solr に HTTP で検索リクエストを送ると、検索結果は XML/JSON で返ってきます。そのため、XML/JSON を HTML などに変換するフロントエンドが

必要ですが Chapter 5 はさまざまなプログラミング言語でフロントエンドをプログラミングする方法を紹介しています。

Chapter 6 以降はさらに Solr を活用したい人たち向けのパートです。Chapter 6 ではリレーショナルデータベース、ファイルサーバなどのさまざまな情報リソースから Solr のインデックスにデータを取り込むためのしくみ「データインポートハンドラ」や「ManifoldCF」の使い方を紹介しています。

Chapter 7 では Solr の検索機能を応用してレコメンデーションやスペルチェックなどの楽しい仕掛けをアプリケーションに付加できる「サーチコンポーネント」の紹介や、全文検索以外の応用的な検索機能である位置情報検索や Join 検索を紹介しています。

Chapter 8 はスケーラビリティを確保するためのしくみである、巨大なインデックスを分割して検索する「分散検索」の設定方法や使い方、レプリケーションの方法、また、Solr 4 から新たに導入された SolrCloud と呼ばれるクラスタ構成の設定方法や使い方についても紹介します。

そして Chapter 9 はユーザ企業の立場から経験上得られた貴重な知見や Solr の癖などの情報を紹介しています。

最後に、付録として各種 API (スキーマ操作、マルチコア操作、コレクション操作) の簡単な説明を記載してあります。

## サンプルコードのダウンロードと本書のサポート

本書で紹介しているサンプルプログラムや設定ファイルは、技術評論社のホームページをはじめ、執筆陣が勤務する下記の会社のホームページからダウンロードできます。

株式会社シーマーク	<a href="http://www.seamark.co.jp">http://www.seamark.co.jp</a>
株式会社ロンウィット	<a href="http://www.rondhuit.co.jp">http://www.rondhuit.co.jp</a>

本書の執筆には万全を期しましたが、ソフトウェア同様、残念ながら不具合が紛れ込む可能性があります。万一誤りを発見したり、手順どおり試したのに動かないことなどがありましたら、上記ホームページの問い合わせフォームなどからお気軽にご連絡ください。また、下記ブログのコメント欄でも受け付けます。

@johtani の日記 2nd <http://blog.johtani.info>

いただいた問い合わせには、個別に回答したり、ブログの記事上で回答したり、正誤表を作成したりなど、何らかの対応をしたいと思えます。なお対応には時間がかかる場合がありますので、あらかじめご了承ください。

# Contents

[ 目次 ]

はじめに	iii
謝辞	iv
対象読者	iv
本書の構成と読み方	iv
サンプルコードのダウンロードと本書のサポート	v

## Chapter 1

### イントロダクション 001

<b>1.1</b> Apache Solrとは何か?	002
1.1.1 Solrの特徴	003
<b>1.2</b> 全文検索と転置索引(インデックス)の基礎知識	006
<b>1.3</b> 形態素解析と N-gram	008
<b>1.4</b> Solrのインストール	009
1.4.1 Javaのインストール	009
1.4.2 Solrのインストール(全プラットフォーム共通)	010
<b>1.5</b> Solrの付属サンプルを実行する	011
1.5.1 Solrの起動	011
1.5.2 サンプルデータの登録	012
1.5.3 サンプルデータの検索	014
<b>1.6</b> Solrで日本語を扱う	015
1.6.1 N-gram	015
1.6.2 形態素解析	016
<b>1.7</b> Solrのアーキテクチャ	017
<b>1.8</b> Solrホームディレクトリ	018
1.8.1 ディレクトリ構成	018
<b>1.9</b> Solrの使い方	020

## Chapter 2

### スキーマの設定 021

<b>2.1</b> スキーマ定義ファイル	022
2.1.1 schema.xmlの構成要素	022
2.1.2 schema.xmlファイルの配置場所	023
2.1.3 schema.xmlのバージョン	023
Column ~バージョン1.5のSchema差異について~	024
<b>2.2</b> フィールド型	024
2.2.1 非テキスト系フィールド型	026
2.2.2 テキスト系フィールド型	029
Column ~ autoGeneratePhraseQueries属性~	031
2.2.3 文字フィルタ	031
2.2.4 代表的なトークナイザ	033
Column ~新NGramTokenizerFactory~	035
2.2.5 代表的なトークンフィルタ	035
Column ~文字フィルタとトークンフィルタ~	038
<b>2.3</b> フィールドの定義	039
2.3.1 フィールドの定義	039
2.3.2 ダイナミックフィールドの定義	041
2.3.3 ユニークキーフィールドの定義	041
Column ~デフォルト検索フィールドと、デフォルトオペレータの定義について~	042
2.3.4 コピーフィールドの定義	042
Column ~特殊なフィールドと設定~	043
<b>2.4</b> Similarityの定義	043
<b>2.5</b> 書籍サンプルでの定義例	044
2.5.1 書籍データフィールドの説明	046
2.5.2 フィールド型text_ja	048
2.5.3 フィールド型text_cjk	048
<b>2.6</b> analysis.jspを使う	049

2.6.1	text_ja	049
2.6.2	text_cjk	051

## Chapter 3

<b>インデックスの作成</b>		<b>053</b>
<b>3.1</b>	ドキュメントの追加	054
3.1.1	XMLファイルによる登録	054
3.1.2	JSONファイルによる登録	057
3.1.3	CSVファイルによる登録	059
3.1.4	stream.file/stream.urlパラメータによる登録	060
3.1.5	インデックスディレクトリ	061
<b>3.2</b>	ドキュメントの更新	062
3.2.1	Num Docs と Max Doc	062
3.2.2	<add/>の overwrite 属性	064
3.2.3	アトミックアップデート	064
3.2.4	バージョンについて	066
<b>3.3</b>	ドキュメントの削除	068
<b>3.4</b>	コミット/最適化/ロールバック	069
3.4.1	コミット	070
3.4.2	最適化	071
3.4.3	自動コミット	071
3.4.4	commitWithin	072
3.4.5	ロールバック	072
3.4.6	管理画面の Stats	073
<b>3.5</b>	バッチ処理	073
<b>3.6</b>	インデックス作成に関連する solrconfig.xml の設定	074
<b>3.7</b>	インデックスの内容を見る	075
3.7.1	スキーマブラウザ	075

## Chapter 4

<b>検索する</b>		<b>077</b>
<b>4.1</b>	検索の基本動作	078
4.1.1	検索のしくみ	078
4.1.2	検索リクエストとパラメータ	079
4.1.3	検索式	082
4.1.4	ファンクションクエリ	086
4.1.5	リクエストハンドラ(サーチハンドラ)	087
4.1.6	レスポンスライタ	088
4.1.7	リアルタイムGET	089
<b>4.2</b>	検索結果レスポンス	090
4.2.1	ヘッダ情報(responseHeader)	091
4.2.2	検索結果(result)	091
<b>4.3</b>	結果のソート	092
4.3.1	スコアによるソート	092
4.3.2	特定フィールドでのソート	093
4.3.3	FunctionQueryでのソート(Solr 3.1以降)	093
4.3.4	ランダムソート	094
<b>4.4</b>	ハイライト(要約と検索語の強調表示)	095
4.4.1	ハイライトの基本	096
4.4.2	ハイライトの結果(XML)	097
4.4.3	その他ハイライト用のパラメータ	097
4.4.4	ハイライタの種類	098
4.4.5	ハイライタの設定	099
Column	~ハイライト処理を速くするには~	101
<b>4.5</b>	ファセットの表示と絞り込み検索	101
4.5.1	ファセットとは	101
4.5.2	Solr のファセット機能	103
4.5.3	絞り込み検索(フィルタクエリ:fq)	110
4.5.4	ファセットのしくみ	111

<b>4.6</b> キャッシュ	112
4.6.1 各種キャッシュ	112
4.6.2 キャッシュの設定	114
4.6.3 キャッシュの初期生成および再生成	114
4.6.4 キャッシュの統計情報	115

## Chapter 5

### クライアントプログラミング 117

<b>5.1</b> Java	120
5.1.1 SolrJの構成	120
5.1.2 インデクシングプログラム	122
5.1.3 SolrJによる検索プログラム	125
<b>5.2</b> PHP	130
5.2.1 solr-php-clientのインストール	130
5.2.2 solr-php-clientによる検索プログラム	130
<b>5.3</b> Ruby	137
5.3.1 RSolrのインストール	137
5.3.2 RSolrによる検索プログラム	138
<b>5.4</b> Python	144
5.4.1 Pysolrのインストール	144
5.4.2 Pysolrによる検索プログラム	145
<b>5.5</b> Perl	151
5.5.1 Webservice-Solrのインストール	151
5.5.2 Webservice-Solrによる検索プログラム	152

## Chapter 6

### データのクローリング 159

<b>6.1</b> クローリングの全体像	160
6.1.1 クロールとは	161

6.1.2 インデクシングの前処理	162
6.1.3 データインポートハンドラ(DIH)によるインデクシング	162
6.1.4 ManifoldCFによるインデクシング	162
6.1.5 DIHとManifoldCFの機能比較	162
<b>6.2</b> インデクシングの前処理——UpdateRequestProcessor	163
6.2.1 言語判別処理 ——LangDetectLanguageIdentifierUpdateProcessorFactory	164
6.2.2 スクリプト処理——StatelessScriptUpdateProcessorFactory	167
6.2.3 その他のUpdateRequestProcessor	167
<b>6.3</b> データインポートハンドラ(DIH)	168
6.3.1 データインポートハンドラの設定方法	168
6.3.2 RDBからのインポート	171
6.3.3 XMLからのインポート	175
<b>6.4</b> ManifoldCF	178
6.4.1 ManifoldCFの概要とインストール	178
6.4.2 ManifoldCFを使う(ローカルファイルシステムのクロール)	182
6.4.3 ManifoldCFとの連携	187
6.4.4 ManifoldCFを使う(Windowsサーバのクロール)	190

## Chapter 7

### より高度な検索 197

<b>7.1</b> サーチコンポーネントとサーチハンドラ	198
7.1.1 検索語のサジェスション——TermsComponent	200
7.1.2 「もしかして……」の実現——SpellCheckComponent	203
7.1.3 Suggester	208
7.1.4 お勧め商品の提示——MoreLikeThisComponent	214
7.1.5 相場の表示——StatsComponent	215
7.1.6 意図的なランクアップ——QueryElevationComponent	217
7.1.7 ドキュメントの単語情報の表示——TermVectorComponent	219

7.1.8	検索結果のクラスタリング — ClusteringComponent	222
<b>7.2</b>	<b>グルーピング検索 — Result Grouping/Field Collapsing</b>	225
7.2.1	グルーピング検索の利用	225
7.2.2	グルーピング検索のサンプル	226
<b>7.3</b>	<b>空間検索 — Spatial Search</b>	229
7.3.1	Spatialとは	229
7.3.2	事前準備	230
Column	～緯度・経度の指定方法～	231
7.3.3	空間検索の利用	231
Column	～算出された距離の利用～	231
<b>7.4</b>	<b>Join 検索</b>	233
7.4.1	Join 検索の利用	234
7.4.2	Join 検索のサンプル	234

## Chapter 8

### クラスタ構築と運用 237

<b>8.1</b>	<b>分散インデックス/分散検索</b>	238
8.1.1	分散インデックス/分散検索とは	239
8.1.2	分散インデックス/分散検索環境のセットアップ	240
8.1.3	分散インデクシングの実行	242
8.1.4	分散検索の実行	243
8.1.5	分散インデックス/分散検索の運用上の注意	245
<b>8.2</b>	<b>レプリケーション</b>	249
8.2.1	Solrのレプリケーション	250
8.2.2	マスタ/スレーブの設定	252
8.2.3	マスタ/スレーブのセットアップ	253
8.2.4	レプリケーションの実行	257
8.2.5	リピーター	261
<b>8.3</b>	<b>分散インデックス, 分散検索, レプリケーションを利用したクラスタ</b>	262

<b>8.4</b>	<b>SolrCloud</b>	263
8.4.1	SolrCloudパラメータ	265
8.4.2	SolrCloud環境のセットアップ	266
8.4.3	SolrCloudの起動	267
8.4.4	設定の中央集中管理	269
8.4.5	耐久性のある書き込み	272
8.4.6	自動フェールオーバー	274
8.4.7	リーダー選出	275
8.4.8	コマンドラインインターフェース	277
8.4.9	用語集	278

## Chapter 9

### 開発および運用のTIPS 279

<b>9.1</b>	<b>設計</b>	280
9.1.1	サイジング	280
9.1.2	シノニム	281
9.1.3	ユーザ辞書	285
9.1.4	ポイントとなるSolrの動き	286
<b>9.2</b>	<b>モニタリング</b>	290
9.2.1	管理画面からの確認	290
9.2.2	REST APIからの確認	290
9.2.3	JMXからの確認	292
<b>9.3</b>	<b>チューニング</b>	296
9.3.1	パフォーマンスチューニング	296
9.3.2	精度チューニング	298
<b>9.4</b>	<b>Solr 1.4系からのバージョンアップ</b>	303
9.4.1	バージョンアップ時の注意点	303
<b>9.5</b>	<b>運用</b>	304

## 付録

<b>Appendix 1</b>	<b>スキーマ操作</b>	<b>306</b>
<b>A.1.1</b>	フィールド情報の取得	306
<b>A.1.2</b>	フィールドの追加	307
<b>Appendix 2</b>	<b>マルチコア操作</b>	<b>309</b>
<b>A.2.1</b>	操作方法	309
<b>A.2.2</b>	STATUS	310
<b>A.2.3</b>	CREATE	310
<b>A.2.4</b>	RELOAD	311
<b>A.2.5</b>	RENAME	312
<b>A.2.6</b>	SWAP	312
<b>A.2.7</b>	UNLOAD	314
<b>A.2.8</b>	MERGEINDEXES	314
<b>A.2.9</b>	SPLIT	315
<b>Appendix 3</b>	<b>コレクション操作</b>	<b>317</b>
<b>A.3.1</b>	操作方法	317
<b>A.3.2</b>	CREATE	317
<b>A.3.3</b>	DELETE	320
<b>A.3.4</b>	RELOAD	320
<b>A.3.5</b>	SPLITSHARD	321
<b>A.3.6</b>	CREATEALIAS	323
<b>A.3.7</b>	DELETEALIAS	324
<b>A.3.8</b>	CoreAdmin APIを使ったコアの作成	325
	索引	327