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 1

1.1 Apache Solrとは 2
1.1.1 Solr の特徴 3

1.2 検索エンジンのしくみ 5
1.2.1 転置インデックス 5
1.2.2 検索結果のランキング 7

1.3 単語分割とアナライザ 10
1.3.1 単語分割（トークナイズ） 10
1.3.2 アナライザ 10

1.4 Solrのインストール 11
1.4.1 Java のインストール 11
1.4.2 Solr のインストール（全プラットフォーム共通） 12

1.5 Solr付属のサンプルを実行する 13
1.5.1 Solr の起動とサンプルデータの登録 13
1.5.2 サンプルデータの検索 14
1.5.3 Solritas の使い方 15
1.5.4 Solr の停止 17

1.6 Solrのアーキテクチャ 17
1.6.1 Solr 全体構成 17
1.6.2 Solr ホームディレクトリ 18

Chapter 2 スキーマの定義 21

2.1 スキーマとは 22

2.2 スキーマ定義ファイル 22
2.2.1 コアの作成 22
2.2.2 managed-schema と schema.xml 24
2.2.3 スキーマ定義ファイルの配置場所 24

2.3 スキーマ定義の流れ 25
2.3.1 スキーマ定義の有効化 26

2.4 フィールドタイプ 26
2.4.1 フィールドタイプの定義 27
2.4.2 非テキスト系フィールドタイプ 28
2.4.3 テキスト系フィールドタイプ 29
2.4.4 アナライザ 30

2.5 フィールドの定義 32

2.6 Analysis画面 35
2.6.1 text_ja の動作確認 36

2.7 書籍データでの定義例 39
2.7.1 書籍データのフィールドについて 40

Chapter 3 インデックスの作成 45

3.1 ドキュメントの登録 46
3.1.1 Solr の起動 46
3.1.2 コアの作成 46
3.1.3 スキーマの定義 46
3.1.4 JSON ファイルによる登録 47
3.1.5 XML ファイルによる登録 50
3.1.6 CSV ファイルによる登録 52
Column stream.file／stream.url パラメータによる登録 54

3.2 「書籍データ」のインデクシング 54
3.2.1 sample-books.json の登録 55
3.2.2 ドキュメントの登録確認 55

3.3 インデックスディレクトリ 57

3.4 ドキュメントの更新 57
3.4.1 アトミックアップデート（部分更新） 58

3.5 ドキュメントの削除 61

3.6 コミット／ロールバック 62
3.6.1 コミット 62
3.6.2 自動コミット 63
3.6.3 commitWithin 64
3.6.4 ロールバック 65

Chapter 4 ドキュメントの検索 67

4.1 検索の基本動作 68
4.1.1 検索のしくみ 68
4.1.2 検索リクエストとパラメータ 69
4.1.3 検索式 72
4.1.4 サーチハンドラ 73
4.1.5 レスポンスライタ 74

4.2 検索レスポンス 74
4.2.1 ヘッダ情報 75
4.2.2 検索結果 76

4.3 検索結果のソート 77
4.3.1 スコアでのソート 77
4.3.2 特定フィールドでのソート 77

4.4 ハイライト 78
4.4.1 ハイライトの基本動作 78
4.4.2 ハイライトの設定 79
4.4.3 ハイライトの結果 79

4.5 絞り込み検索（フィルタクエリ：fq） 80

4.6 ファセット 81

4.6.1 ファセットとは 81
4.6.2 ファセットのしくみ 84

Chapter 5 高度なインデクシング 87

5.1 バッチ処理 88

5.1.1 JSON データによるバッチ処理例 88
5.1.2 XML データによるバッチ処理例 89

5.2 インデクシング前処理——UpdateRequestProcessor 90

5.2.1 UpdateRequestProcessor の設定 90
5.2.2 正規表現による置換——RegexReplaceProcessorFactory 91
5.2.3 スクリプト処理——StatelessScriptUpdateProcessorFactory 93
5.2.4 その他の UpdateRequestProcessor 95

5.3 データのインポート——DataImportHandler 96

5.3.1 データインポートハンドラの設定 96
5.3.2 RDB からのインポート 98

5.4 擬似リアルタイム検索 104

5.4.1 （ハード）コミットとソフトコミット 104
5.4.2 自動ソフトコミット 104
Column 管理画面の Stats 105

5.5 バイナリ形式ドキュメントのインデクシング——ExtractingRequestHandler 105

5.5.1 ExtractingRequestHandler の設定 106
5.5.2 ExtractingRequestHandler を使ったインデクシング 106

5.6 インデックス作成に関連する solrconfig.xml の設定 109

5.6.1 IndexConfig 設定 109
5.6.2 インデックスセグメントとマージポリシー 110
5.6.3 オプティマイズ 111

5.7 トランザクションログ 113

5.7.1 トランザクションログとインデックスの更新のしくみ 113
5.7.2 トランザクションログの設定 114
5.7.3 リアルタイム Get 115
5.7.4 インデックスのリカバリ 118

Chapter 6 高度な検索 119

6.1 クエリパーサ 120

6.1.1 Standard クエリパーサ 120
6.1.2 DisMax クエリパーサ 122
6.1.3 Extended DisMax クエリパーサ 123
6.1.4 ローカルパラメータ 124

6.2 ハイライト 125

6.2.1 ハイライタの種類 127
6.2.2 ハイライタの設定 127
6.2.3 ハイライト用の検索パラメータ 129

6.3 ファセット 131

6.3.1 フィールド値によるファセット 132
6.3.2 クエリによるファセット 133
6.3.3 レンジファセット（範囲によるファセット） 134
6.3.4 ピボットファセット 136

6.4 サーチコンポーネントとサーチハンドラ 138

6.4.1 検索キーワードのサジェスチョン——SuggestComponent 141
6.4.2 統計情報の表示——StatsComponent 146

6.5 Result Grouping と Collapse and Expand 150

6.5.1 Result Grouping とは 150
6.5.2 Collapse and Expand とは 155

6.6 空間検索 159

6.6.1 空間検索とは 159
6.6.2 フィールド定義とインデックス 160
Column 地点情報のデータ形式 161

Column geodist() の計算の誤差 162
6.6.3 空間検索の利用例 163

6.7 ファンクションクエリ 166
6.7.1 ファンクションクエリとは 166
6.7.2 Solr に標準のファンクション 169

6.8 キャッシュ 172
6.8.1 キャッシュの種類と設定 172
6.8.2 キャッシュの自動ウォームアップ 174
6.8.3 キャッシュの統計情報 175

Chapter 7 スキーマ設計 177

7.1 スキーマ定義ファイル 178
7.1.1 スキーマ定義ファイルの配置場所と設定 178
7.1.2 スキーマ定義の構成要素 179
7.1.3 スキーマレスモード 179

7.2 フィールドタイプ 180
7.2.1 フィールドタイプの設定オプション 180
7.2.2 非テキスト系フィールドタイプ 184
7.2.3 テキスト系フィールドタイプ 185
7.2.4 アナライザ 186
7.2.5 文字フィルタ 188
7.2.6 代表的なトークナイザ 190
7.2.7 代表的なトークンフィルタ 192

7.3 フィールドの定義 196
7.3.1 フィールドの定義 197
7.3.2 ダイナミックフィールドの定義 198
7.3.3 ユニークキーフィールドの定義 199
7.3.4 コピーフィールドの定義 199
7.3.5 その他のフィールドとオプションの説明 200

7.4 Similarityの定義 201

Chapter 8 クラスタ構築と運用 203

8.1 単一ノードの限界 204

8.2 分散インデックスと分散検索 205
8.2.1 分散インデックス 205
8.2.2 分散検索 205
8.2.3 分散インデックスのセットアップ 206
8.2.4 分散インデクシングと分散検索の実行 207
8.2.5 分散検索のエラー回避 209

8.3 レプリケーション 211
8.3.1 レプリケーションの概要 211
8.3.2 マスタ／スレーブの設定方法と設定項目 212
8.3.3 マスタ／スレーブのセットアップ 214
8.3.4 レプリケーションの確認 216

8.4 レガシーなクラスタ 219
8.4.1 レガシーなクラスタを構築する準備 220
8.4.2 レガシーなクラスタの構築 220
8.4.3 分散インデクシングと分散検索 222

8.5 SolrCloud 223
8.5.1 ZooKeeper のインストールと起動 224
8.5.2 SolrCloud 構築 225
8.5.3 分散インデクシングと分散検索，レプリケーション 231
8.5.4 フェールオーバー 232
8.5.5 インデックスのリカバリ 235
8.5.6 リーダーの選出 238
8.5.7 SolrCloud の拡張 239

Chapter 9 検索精度の改善 245

9.1 検索精度の定義 246
9.1.1 再現率と適合率 246
9.1.2 ランキング 248

9.2 再現率と適合率の改善 249

9.2.1 アナライザの変更による再現率と適合率の改善 249
9.2.2 各種辞書の活用 264

9.3 ランキングの改善 269

9.3.1 キーワードと文書の類似度 269
9.3.2 ファンクションクエリの活用 273
9.3.3 クエリリランキング 275

9.4 検索精度の評価 277

9.4.1 オフライン評価 278
9.4.2 オンライン評価とオフライン評価 284
9.4.3 A/B テスト 286

9.5 機械学習による検索の改善 291

9.5.1 Learning To Rank 292

9.6 Solrをレコメンドエンジンとして使う 296

9.6.1 レコメンドは検索の特殊ケース 296
9.6.2 レコメンドエンジンの作り方の例 297
9.6.3 ログからレコメンドの中身を考える 300
9.6.4 クエリに任せる処理，インデックス時に行う処理 302
9.6.5 ID と属性の比較 303

Chapter 10 開発運用の TIPS 305

10.1 サイジング 306

10.1.1 CPU 306
10.1.2 メモリ 306
10.1.3 ストレージ 307

10.2 モニタリング 307

10.2.1 管理画面からの確認 307
10.2.2 REST API からの確認 308
10.2.3 JMX からの確認 309

10.3 Solr 4系からのバージョンアップ 312

10.3.1 Java のバージョンの変更 312

10.3.2 war デプロイの廃止および起動方法の変更 312
10.3.3 solr.xml フォーマットの変更 313
10.3.4 デフォルトの SchemaFactory の変更 313
10.3.5 デフォルトの Similarity の変更 313
10.3.6 StopFilterFactory の enablePositionIncrements 属性の廃止 313
10.3.7 SolrJ API の変更 313
10.3.8 DocValues の使用 314

10.4 ログの設定 314

10.4.1 Solr ログ 314
10.4.2 GC ログ 317

10.5 Distributed IDF 317

10.6 FAQ集 318

10.6.1 Question 1 － ソートがうまくいかない 318
10.6.2 Question 2 － シノニム辞書を定義したのに思うように動作しない 318
10.6.3 Question 3 － 定義できるフィールド数に制限はあるのか？ 320
10.6.4 Question 4 － ワイルドカード検索がうまくいかない 321
10.6.5 Question 5 － Solr の起動が極端に遅い 321
10.6.6 Question 6 － 実際にレコメンド目的で Solr を導入している例とは？ 322

Chapter 11 SolrJ プログラミング 325

11.1 SolrJクライアントアプリケーション 326

11.1.1 ビルドとインストール 326
11.1.2 ドキュメントの登録 328
11.1.3 ドキュメントの検索 333
11.1.4 ドキュメントの削除 337
11.1.5 インテグレーションテスト（結合テスト） 338

Appendix 付録 343

Appendix 1 Gitのインストール 344

A.1.1 Ubuntu などの Debian 系 Linux 344
A.1.2 CentOS など RedHat 系 Linux 344
A.1.3 Mac／Windows 344

Appendix 2 Python機械学習の環境構築 346

A.2.1 必要なもの 346
A.2.2 インストール 346
A.2.3 関連パッケージの準備 346
A.2.4 Jupyter Notebookの起動 347

Appendix 3 ネステッドドキュメントとBlock Join 348

A.3.1 ユースケース 348
A.3.2 インデクシング 348
Column Luceneインデックスとネステッドドキュメント 349
A.3.3 Block Join 349
A.3.4 ネステッドドキュメントにおけるファセット 353

Appendix 4 N-best 356

A.4.1 形態素解析のしくみ 356
A.4.2 N-bestによって改善できること 358
A.4.3 N-bestの仕様 359
A.4.4 N-bestのパラメータ 360
A.4.5 Analysis画面での確認 360

索引 363