Railsアプリケーション開発者のためのSolr入門

Solr 1.2対応版

株式会社 ロンウイット

rel 1.0

RONDHUIT

WWW.RONDHUIT.COM

はじめに

RONDHUIT

- このたびは無料セミナー「Railsアプリケーション開発者のためのSolr入門」にご参加くださいまして、まことにありがとうございます。

- 本書はApache Solr 1.2の基本的な使い方について説明しています。

- 本書の権利は株式会社ロンウイットが保有しています。

Goal/Non-Goal

RONDHUIT

- Goal
  - Apache Solr 1.2（以下Solr）をインストールし、起動できるようになる
  - Solrの管理画面の基本的な使い方を理解する
  - Solrへの文書登録の基本を理解する
  - Solrを使って基本的な検索ができるようになる
  - Solrを使ったファセットカウントの取得～絞り込み検索の流れを理解する
- Non-Goal
  - Solrのすべての機能を理解する
  - Solrの設定ファイルの項目の意味を理解し、使いこなせるようになる
  - Solrの各種プラグインを開発し、Solrを自在にカスタマイズする
  - パフォーマンスチューニングやメモリチューニングができるようになる

内容

RONDHUIT

1. 日本語全文検索の基礎
   - 全文検索の方式
   - 転置索引の理解
2. Solrのインストールと起動
   - Solrのインストール
   - Solrの起動
   - Solr管理画面の起動
3. 付属exampleの使い方
   - exampleデータの登録
   - exampleデータの検索
   - 検索リクエストの基本パラメータ
   - solr-rubyの利用
4. 日本語データの登録と検索
   - schema.xmlの変更
   - 管理画面のANALYSIS機能の利用
5. Rails検索アプリケーション
   - ファセットと絞り込みのサンプルRails検索アプリケーション

日本語全文検索の基礎

RONDHUIT

WWW.RONDHUIT.COM

全文検索の方式

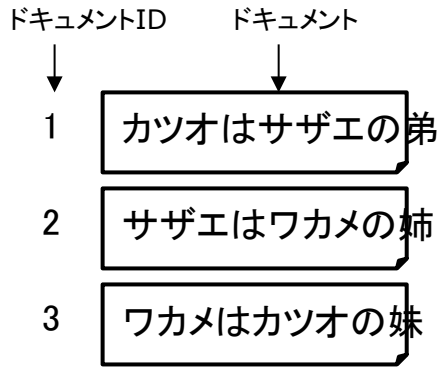
RONDHUIT

- 順次検索方式
  - インデックスを作らない
  - ドキュメントの先頭から、検索質問語の文字列と順次比較する
  - 例：UNIXコマンドのgrep、SQLのlike検索

- 転置索引方式
  - あらかじめ検索対象のドキュメントからインデックスを作成
  - 例：Solr、Namazu、Senna、Google、Yahoo!、・・・

転置索引の作成方法

(a)



(b)

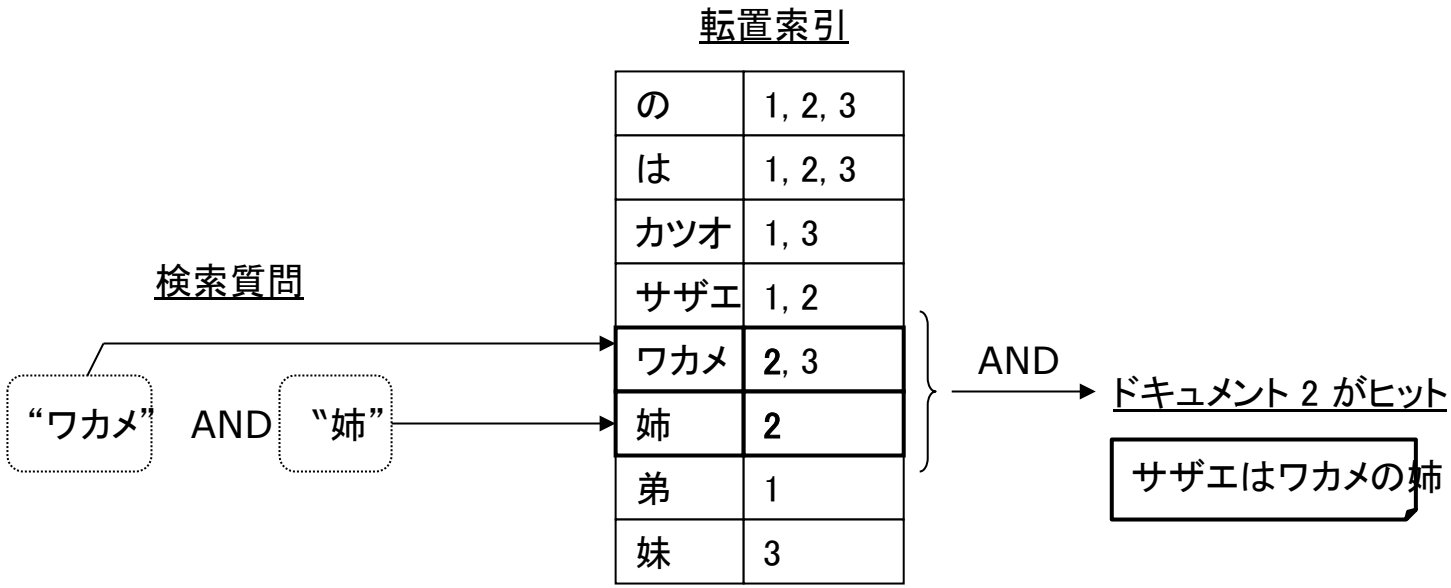
カツオ:1、は:1、サザエ:1、の:1、弟:1、
サザエ:2、は:2、ワカメ:2、の:2、姉:2、
ワカメ:3、は:3、カツオ:3、の:3、妹:3

(c)

の	1, 2, 3
は	1, 2, 3
カツオ	1, 3
サザエ	1, 2

ワカメ	2, 3
姉	2
弟	1
妹	3

転置索引の検索方法



全文検索の各方式の特徴

RONDHUIT

方式	長所	短所	適用例
順次検索方式	•インデックスを使わないので余計なメンテナンス作業が不要 •インデックスを使わないので「今あるドキュメント」の内容をリアルタイムに検索できる	•大量のドキュメントの検索には不向き •多数のユーザから繰り返し検索される状況下ではかなり非効率	•UNIXのgrepコマンド •SQLのlike検索
転置索引方式	•大量ドキュメントを保有し、多数のユーザから繰り返し検索される状況下でも効率的に処理できる •大規模な検索に向く	•インデックスをメンテナンスしなければならないため、「今あるドキュメント」とインデックスの内容に差異が生じる場合がある •インデックスのサイズが巨大になる	•多数のユーザから利用されるアプリケーション •インターネットやイントラネットなどのコンテンツ検索機能

（日本語）全文検索システムの問題

RONDHUIT

- インデックスのサイズ
- インデックスの作成にかかる時間
- ヒットしすぎる
  - 検索結果一覧の表示順序（ランキング）のスコアを計算
  - Google ⇒ PageRank
  - Lucene ⇒ tf*idf（*Similarity*抽象クラスのJavadoc参照）
- 日本語テキスト処理
  - 英語などと違って、単語を識別するのが困難
  - 形態素解析
    - 辞書を用いる方式が主流 ⇒ 流行語に弱い
    - 辞書にない単語が検索されにくい（「東京」で「東京大学」がヒットしない）
    - JapaneseAnalyzer
  - N-gram
    - Nの大きさによりノイズや検索漏れ
    - CJKAnalyzer（「東京都」⇒「東京」「京都」）
  - 表記の揺れ
    - 「インタフェース」「インターフェイス」、「引っ越し」「引越し」
    - 半角・全角、新旧漢字（「亜」と「亞」）

Solrのインストールと起動

（演習）Solrのインストール～起動

RONDHUIT

- インストール
  - apache-solr-1.2.0.zipを展開します（Subversionから作成したパッケージ）
  - これ以降、展開したディレクトリを$SOLRとして参照します
- 起動
  - $SOLR/exampleに移動します
  - start.jarを使って起動します

```
$ cd $SOLR/example
$ java -jar start.jar
2008-06-03 17:20:52.050::INFO:  Logging to STDERR via org.mortbay.log.StdErrLog
2008-06-03 17:20:52.376::INFO:  jetty-6.1.3
                :
                :
2008/06/03 17:20:55 org.apache.solr.servlet.SolrDispatchFilter init
情報: SolrDispatchFilter.init() done
2008/06/03 17:20:55 org.apache.solr.servlet.SolrServlet init
情報: SolrServlet.init()
2008/06/03 17:20:55 org.apache.solr.servlet.SolrServlet init
情報: SolrServlet.init() done
2008/06/03 17:20:55 org.apache.solr.servlet.SolrUpdateServlet init
情報: SolrUpdateServlet.init() done
2008-06-03 17:20:55.675::INFO:  Started SocketConnector @ 0.0.0.0:8983
```

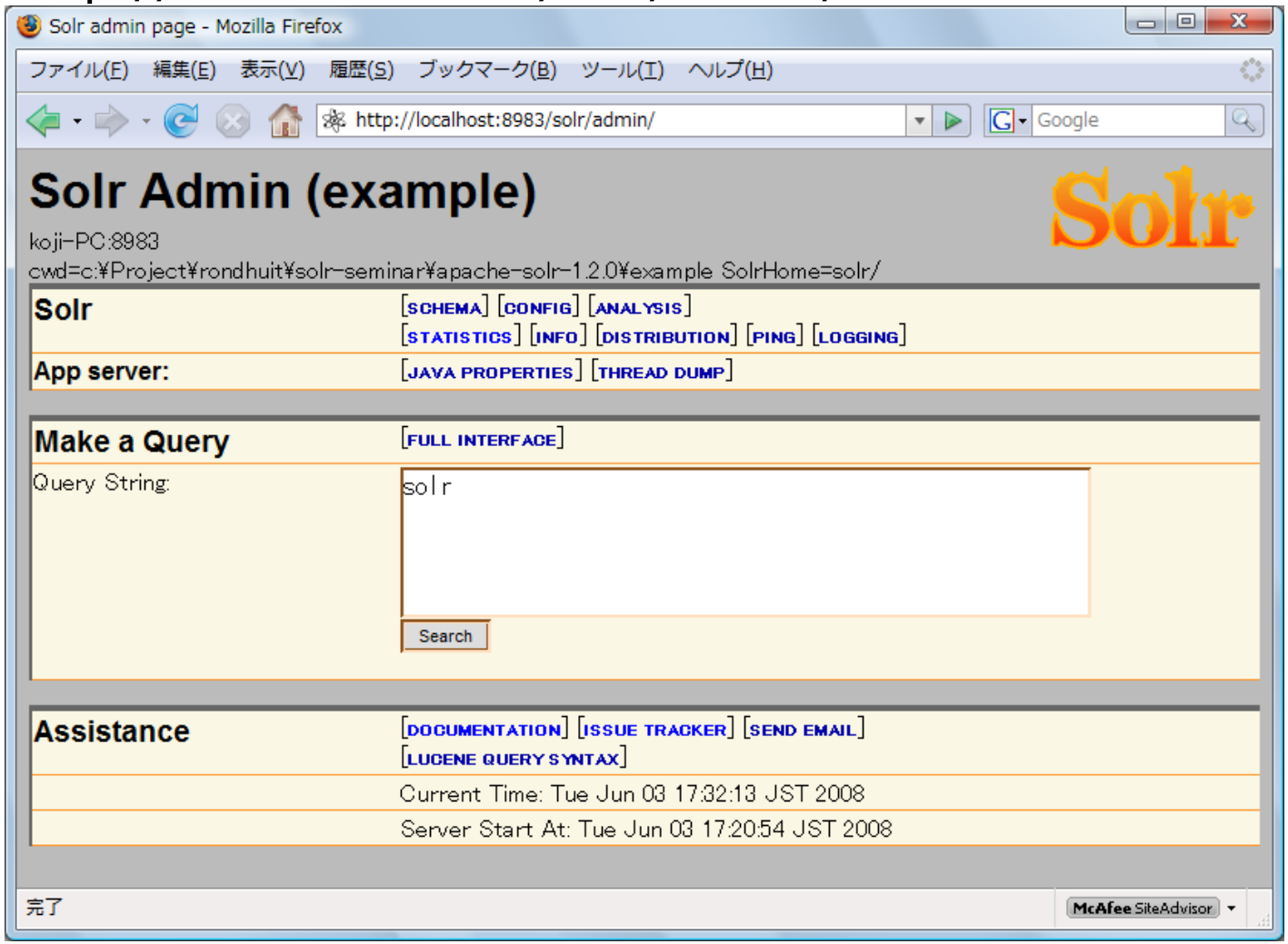
  - 停止は[Ctrl]+[c]

（演習）Solr管理画面の起動

RONDHUIT

- ブラウザを使って管理画面を起動します

  http://localhost:8983/solr/admin/



付属exampleの使い方

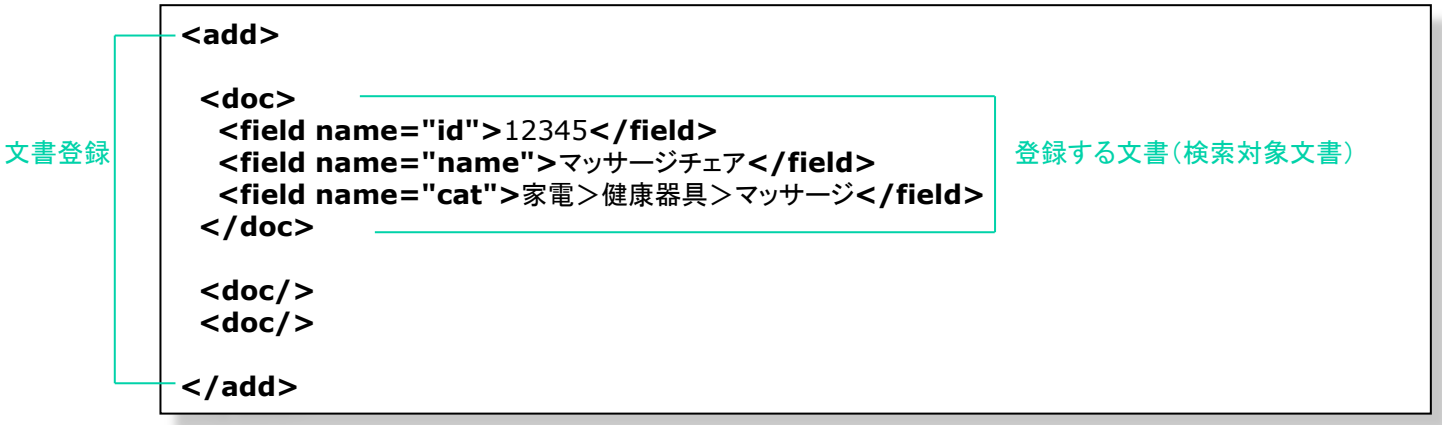
RONDHUIT
WWW.RONDHUIT.COM

（演習）付属exampleデータの登録と検索

- 付属exampleデータの登録
  - $SOLR/example/exampledocsに移動します
  - post.jarを使って登録します

```
$ cd $SOLR/example/exampledocs
$ java -jar post.jar *.xml
SimplePostTool: version 1.2
SimplePostTool: WARNING: Make sure your XML documents are encoded in UTF-8, other
encodings are not currently supported
SimplePostTool: POSTing files to http://localhost:8983/solr/update..
SimplePostTool: POSTing file hd.xml
                :
                :
SimplePostTool: POSTing file vidcard.xml
SimplePostTool: COMMITting Solr index changes..
```

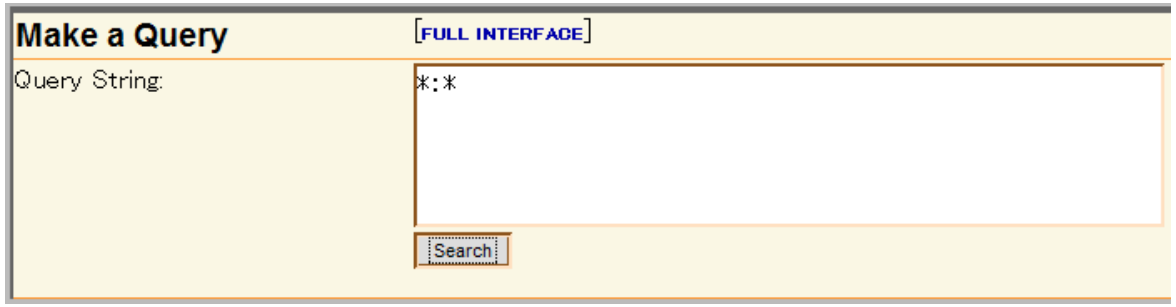
  - 登録されるXMLをエディタで開いて見てみよう




```
<add>

 <doc>
  <field name="id">12345</field>
  <field name="name">マッサージチェア</field>
  <field name="cat">家電＞健康器具＞マッサージ</field>
 </doc>

 <doc/>
 <doc/>

</add>
```

（演習）付属exampleデータの登録と検索

- 付属exampleデータの検索
  - 管理画面から適当な検索語で検索します
    - レスポンスXMLを観察しよう
  - 検索語に"*:*"と指定すると「すべて検索」の意味になります



（演習）検索リクエストの基本パラメータ

RONDHUIT

- 基本パラメータ
  - q＝検索語（検索式）
  - start＝開始文書番号（デフォルト＝0）
  - rows＝ページあたりの文書件数（デフォルト＝10）
  - fl＝フィールド指定
  - wt＝ライタ名（xml（デフォルト）,json,ruby,python,…）
  - indent＝on（出力の字下げを実行）
  - debugQuery＝on（デバッグ出力）
- ファセットカウント取得＆絞り込み検索関連パラメータ
  - facet＝on
  - facet.field＝ファセットフィールド指定
  - facet.mincount＝ファセットフィールド指定時最小カウント指定
  - facet.query＝任意の検索式（範囲検索など）
  - fq＝絞り込み検索条件

http://localhost:8983/solr/select/?q=*%3A*&indent=on&**facet**=on&**facet.field**=cat

（演習）solr-rubyの利用

RONDHUIT

- solr-rubyクライアントライブラリ
  - $SOLR/client/ruby/solr-ruby/lib/solr.rb
- Solrコネクションの取得
  - solr=Solr::Connection.new("http://localhost:8983/solr")
- Solr検索リクエスト
  - request=Solr::Request::Standard.new(:query=>'*:*')
  - response=solr.send(request)
- irbから使ってみよう

```
$ irb -I $SOLR/client/ruby/solr-ruby/lib -r solr

irb(main):001:0> solr=Solr::Connection.new("http://localhost:8983/solr")
=> #<Solr::Connection:0x3b6eb00 @url=#<URI::HTTP:0x1db74d6 URL:http://localhost:
8983/solr>, @connection=#<Net::HTTP localhost:8983 open=false>, @autocommit=false>

irb(main):002:0> request=Solr::Request::Standard.new(:query=>'*:*')
=> #<Solr::Request::Standard:0x3b6b270 @query_type="standard", @params={:query=>
"*:*", :field_list=>["*", "score"]}>

irb(main):003:0> pp solr.send(request)
```

日本語データの登録と検索

RONDHUIT

WWW.RONDHUIT.COM

（演習）日本語データの登録と検索

RONDHUIT

1. インデックスの削除
   - Solrサーバを停止して、$SOLR/example/solr/data/indexディレクトリを削除します

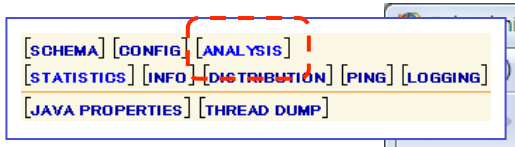
2. schema.xmlの変更
   - $SOLR/example/solr/conf/schema.xmlを入れ替えてSolrサーバを起動します
   - schema.xmlをエディタで開いて見てみよう

3. 日本語サンプルデータの登録
   - post.jar（P.15）を使ってblog.xmlを登録します

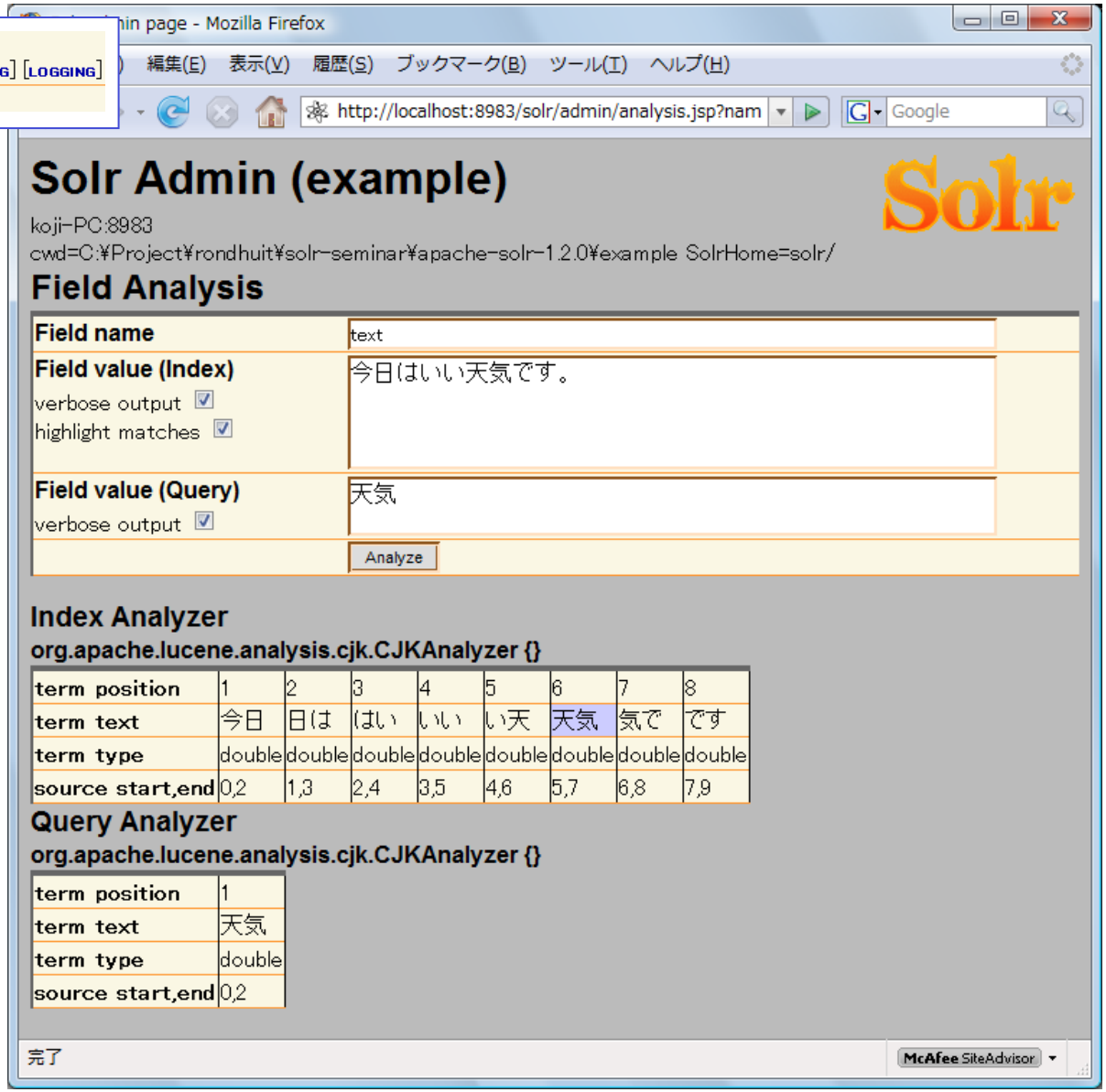
4. 管理画面からの検索
   - 管理画面を起動し、適当な単語で検索してみます
   - （注意）fl=*（デフォルト）を指定した場合、巨大な文書がヒットするとレスポンスが遅くなることがあります

（演習）CJKAnalyzerの動作を理解する

RONDHUIT



管理画面メニューから
[ANALYSIS]を選択



Rails検索アプリケーション

RONDHUIT

WWW.RONDHUIT.COM

（演習）Rails検索アプリケーションの準備

RONDHUIT

1. Railsアプリケーションの雛形生成

```
$ rails sample
```

2. environment.rbの修正
   - ロードパスに"$SOLR/client/ruby/solr-ruby/lib"を加えます
     - 代わりに"$ gem install solr-ruby"でも可

```
$:.unshift '$SOLR/client/ruby/solr-ruby/lib'
```

   - ActiveRecord不使用の設定を行います

```
Rails::Initializer.run do |config|
  config.frameworks -= [ :active_record ]
end
```

3. 既成ファイルのコピー
   - app/controllers/search_controller.rb
   - app/helpers/search_helper.rb
   - app/views/search/*.rhtml
   - lib/solr_context.rb

起動時"`load_missing_constant': uninitialized constant ActiveRecord (NameError)"のエラーになる場合は、config/initializers/new_rails_defaults.rbのActiveRecord定数を参照している2つの行をコメントアウトする。

4. Railsアプリケーションサーバの起動

```
$ ruby script/server
```

（演習）ファセット＆絞り込み検索

RONDHUIT

1. トップページの表示

   http://localhost:3000/search/index

   - しくみを観察しよう
     - qパラメータはどのようになっていますか？
     - ファセットカウント取得をどのように行っていますか？
     - リンクはどのように作成されていますか？

2. 検索結果ページの表示
   - しくみを観察しよう
     - qパラメータはどのようになっていますか？

3. 絞り込み検索の実行
   - しくみを観察しよう
     - fqパラメータはどのようになっていますか？

4. 検索結果[次ページ][前ページ]の表示
   - しくみを観察しよう
     - startパラメータはどのようになっていますか？