Lucene/Solr 3.1 の新機能

2011 年 3 月 22 日
株式会社 ロンウイット　　関口宏司

はじめに

まもなく公開される Lucene/Solr 3.1 の新機能を紹介する。

1. バージョンについて

Solr のユーザーはこれまでのバージョン 1.4 からいきなり 3.1 に番号が飛ぶのに戸惑うかも知れない。これは昨年 3 月に Lucene と Solr のソースコード管理や開発者がマージされたことに伴う副作用のようなものである。Lucene と Solr のソースコード管理ツリーがマージされたことにより、Solr のバージョンが Lucene のそれにあわされた。またこのことから今回のように Lucene と Solr の同時リリースが容易に行えるようになった。（参考：http://www.slideshare.net/KojiSekiguchi/lu-solr-20100709 ）

Lucene は 3.0 から Java 5 以上のサポートとなったが、3.1 から正式に Unicode 4 の補助文字をサポートするようになった。これにより、Analyzer の補助文字に対する動作が 3.0 以前と 3.1 以降では異なることがある。たとえば、補助文字の大文字=>小文字正規化の動作などである。この後方互換性を取るために、Lucene では Version と呼ばれる enum 型を導入した。たとえば、SimpleAnalyzer にて 3.1 以降の動作をさせたい場合は：

```
SimpleAnalyzer a =
 new SimpleAnalyzer(Version.LUCENE_31);
```

という生成の仕方をすることになる。従来の引数なしのコンストラクタは deprecated となり、次のように 3.0 以前の動作となるように設定されている：

```
@Depracated public SimpleAnalyzer(){
    this(Version.LUCENE_30);
}
```

Version 型の引数はほとんどが Analyzer の動作の後方互換性を維持するために使用されているが、例外もある。たとえば QueryParser が PhraseQuery を生成するかどうかを規定する setAutoGeneratePhraseQueries() メソッドのデフォルト設定に影響を与える。日本語では特に N-gram のトークナイザーを利用している場合に注意が必要であろう（参考：http://lucene.jugem.jp/?eid=395 、http://lucene.jugem.jp/?eid=403 ）。Solr ではこの Lucene のデフォルト動作が逆になり、TextField のフィールド型の設定で autoGeneratePhraseQueries="false"を明示しないと Lucene のデフォルトと同じ動作にならない。

Solr における Version 型の設定に関しては、solrconfig.xml の中で次のように行う：

```
<luceneMatchVersion>LUCENE_31
</luceneMatchVersion>
```

また、schema.xml で Lucene の Analyzer を直接使う場合、luceneMatchVersion="LUCENE_31" などと指定する。省略時は LUCENE_24 となる。

2. スキーマや Analyzer

フィールド圧縮がサポートされなくなった（圧縮が

フィールドの仕事ではなくなった)。これにより、Field の byte[]と Store を引数に取るコンストラクターが deprecated になった。これまでの古いインデックスで圧縮されていた Stored フィールドは、マージされるときに自動的に展開されるのでサイズが大きくなる可能性があり、注意が必要である。

Analyzer 関連では、出力するトークン数の上限を抑える LimitTokenCountFilter が追加された。従来インデクシング時にフィールドに登録されるトークン数を抑えるには IndexWriter の setMaxFieldLength()というメソッドを使っていたが、このメソッドは deprecated になり、今後は LimitTokenCountFilter を使うのが推奨となっている。これにより、フィールド毎に上限の設定を変えることが可能となる。そのほかにも、正規表現で文字を正規化する PatternReplaceCharFilter や、ファイルシステム等の階層構造を保持した状態でトークナイズする PathHierarchyTokenizer などが追加されている。

3. SolrJ

SolrJ は Solr サーバーに Java プログラムで書かれたクライアントがアクセスする際に便利に使用できるライブラリである。SolrJ と Solr は javabin と Solr で呼んでいるシリアライズフォーマットを使用しているが、今回このフォーマットが変わり、バージョン 2 となった。従来は、Modified UTF-8 のバイトデータの前に UTF-16 文字数がつけられてやりとりされていたが、これはバグである。3.1 ではこれが修正されて、UTF-8 のバイトデータの前にそのバイト数が正しく送られるようになっている。この変更により、Solr サーバー側を 3.1 にしたら、SolrJ を使用しているクライアント側も同時に 3.1 にする必要がある。

この変更が行われた直後から、バージョンがあわないというエラーが出るのだが、というユーザーからの問い合わせが多く見られるようになった。背景を知っているユーザーでも、クラスパスの設定ミスにより、古い SolrJ を無意識に使ってしまっているとやはりバージョンの不整合が出てしまう。これはなかなか発見が難しい問題なので、現在はエラーメッセージを表示する際に、バージョン番号も表示されるようになっている。

SolrJ はこれ以外にも、クエリファセット時の順序保持などの改善がなされている。

4. 検索一般

64 ビットの Windows または Solaris で JVM が un-map をサポートしている場合（sun.misc.Cleaner クラスと java.nio.DirectByteBuffer クラスの cleaner メソッドが存在していることを指す）、FSDirectory.open()は MMapDirectory を返すようになった。これは検索時にインデックスファイルを仮想メモリ空間にマッピングする Directory 実装である。インデックスサイズと物理メモリ量の見合いで、環境やアプリケーションによっては検索性能を向上させることができるかもしれない。

multiValued=”true”のフィールドで、ソートや関数クエリの適用は許可されないが、これまではフィールド値の内容（ユニークターム数）によっては「成功」しているように動作していた。3.1 以降ははっきりと「失敗」するように動作仕様が変更になった。

検索についてはそれ以外に、次のような改善や機能追加がある：

- 関数クエリの結果によるソート
- 拡張 DisMax クエリパーサーのサポート
- 一般的な範囲ファセットのサポート。これにより、日付型の範囲ファセットが deprecated となった

5. 分散検索

分散検索については、次のような改善があった：

- SpellCheckComponent の分散検索対応
- TermsComponent の分散検索対応
- StatsComponent を分散検索環境でファセット付きで使用しているとき、NPE が出る問題に対応

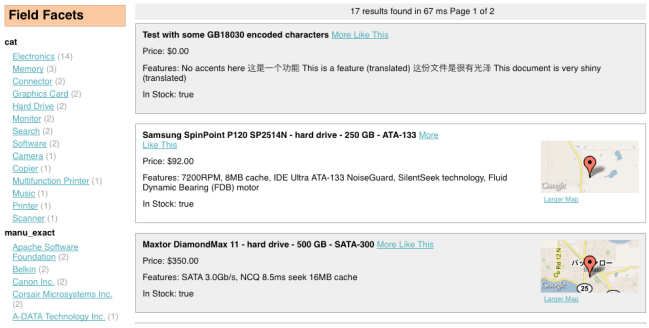
なお、SearchComponent ごとの分散検索対応状況は、http://lucene.jugem.jp/?eid=406 にまとめられている。

6. 空間検索／地理検索

空間検索／地理検索に関しては、次のような改善や機能追加があった：

- Solr の関数 dist（マンハッタン距離、ユークリッド距離）、sqedist（正方形型のユークリッド距離。演算の負担が少ない）、strdist（スペルチェッカーなどに使われる編集距離）の追加
- Solr の関数 hsin（Great Circle Haversine）、ghhsin（GeoHash-hsin）の追加。またこれらの関数を使うにあたって単位の変換を行う geohash(), deg(), rad()の関数の追加
- PointType, LatLonType, GeoHashField の追加

なお、Solr 3.1 の付属サンプルデータには、架空のショップの位置情報が追加され、Solr を起動してサンプルデータを登録し、http://localhost:8983/solr/browse にアクセスすると、下図のような GEO サーチ機能付きの画面が表示されるようになった：



7. ハイライター

Lucene 2.9 から登場した FastVectorHighlighter を Solr 3.1 でサポートした。これにより N-gram フィールドを問題なくハイライト表示できるようになった。solrconfig.xml でのハイライターの設定がこれまでは<config>の直下に書いていたが、この方法は deprecated となり、新しい書式では HighlightComponent の中に書くようになったので注意していただきたい。このほかにも、Lucene の Encoder を Solr で導入したり等、細かな改善がされている。

8. マルチコア

Solr 3.1 では ALIAS コマンドのサポートがなくなった。また、シングルコアの利用時でも solr.xml を Solr ホームディレクトリに配置することが推奨されている。

9. レプリケーション

Java ベースのレプリケーション機能における日時をファイル名の一部に採用しているときの不具合（午後 1 時が 1300 とならずに 0100 となってしまう）の修正や、スクリプトベースのレプリケーションで rsync 稼働時のデータ転送量の調整ができるようにした改善などが加えられている。

なお、マスター／スレーブ構成を採用しているプロジェクトでは、旧バージョンからのアップグレードは、スレーブ側を最初に行うこと。マスター側を最

初にアップグレードしてしまうと、新しいインデックスがスレーブに転送されたとき、スレーブの古い Solr が新しいフォーマットのインデックスを読めないからである。

10. DIH

3.1 では（Solr 1.4 から内蔵された）Apache Tika を利用することで Microsoft のオフィスファイルや PDF ファイル等も DIH のインタフェースを通じてインポートできるようになった。また、マルチスレッドで動作させられるようになった。

さらに、Solr 1.3/1.4 で発見されたさまざまな不具合修正もされている。弊社では DIH の教育トレーニングコース（参考：http://www.rondhuit.com/training.html ）を提供しているが、このコーステキストを準備中にも実にさまざまな不具合に遭遇した。その中には「子エンティティがルートエンティティのときに、差分インポートが正しく動作しない」という重大なものもあったが、3.1 では修正されている（参考：https://issues.apache.org/jira/browse/SOLR-2252 ）。

なお、従来は DIH の JAR ファイルが solr.war に含まれた状態で配布されていたが、他の contrib パッケージ同様 3.1 からは DIH は war に含まれなくなったので、solrconfig.xml の<lib/>設定に DIH を含める必要がある。

11. その他

その他の新機能や改善には、以下のようなものがある：

- Solr の Apache UIMA の取り込み
- Solr に登録する XML ファイルを UTF-8 以外のエンコーディングでも可能にする改善
- Solr に検索専用の RAMDirectory を使えるようにする機能の追加
- Javadoc に google-code-prettify を適用し、XML 設定例やサンプルプログラムをユーザーにとって読みやすくする改善

最後の Javadoc の改善は、弊社からのコミュニティへの提案が反映されたものである。弊社では、Solr の日本語検索機能や集合知利用等の各種機能強化にサポートサービスを加えた「サブスクリプション・パッケージ」を販売している。その機能強化部分の Javadoc は以下で公開している：

http://www.rondhuit-demo.com/RCSS/api/

この Javadoc 内で google-code-prettify を利用しており、JavaScript によるシンタックスハイライティングを行っている（例として、analysis パッケージ内のファクトリクラスを参照されたい）。弊社は Lucene/Solr の Javadoc でもこれを使うべきであるとコミュニティに提案した。するとすぐさま Javadoc 生成方法に関する Ant 実行時のいくつかの改善コメントが寄せられた。弊社ではその改善を自社に逆輸入することによって自社製品のメンテナンス性を改善することができた。OSS とビジネスが融合した今風のエコシステムを実感した瞬間である。・・・もっとも一番偉いのは、prettify を Apache License 2.0 で公開してくれた Google であることはいうまでもない。

（株）ロンウイットについて
ロンウイットはオープンソースの全文検索エンジン Lucene /Solr を企業システムに導入する支援サービス事業を展開している。
お問い合わせ先
〒100-0005 東京都千代田区丸の内 1-8-3
丸の内 トラストタワー本館 20 階
電話：03-5288-5927　　FAX：03-5288-5928
メール：sales@rondhuit.com
ホームページ：http://www.rondhuit.com/