Solr 1.4 の新機能

2009 年 11 月 1 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 Solr 1.4 の新機能を紹介する。

1. Lucene 2.9.1

Solr 1.4 は Lucene 2.9 を使用している。前バージョンの Solr 1.3 は Lucene 2.4-dev であった。2.4-dev からの主な新機能は、IndexReader の取得やソートや範囲検索や TokenStream 再利用による高速化、地理空間検索機能、文字レベル正規化、N-gram を正式にサポートする新しいハイライタ（FastVectorHighlighter）などであるが、詳しくはホワイトペーパー「Lucene 2.9 の新機能」を参照していただきたい。ただしこれらのうち FastVectorHighlighter はスケジュールの関係上、Solr 1.4 には取り込むことができなかった。

なお、Solr 1.4 は正確には Lucene 2.9.0 のバグフィックス版である 2.9.1 を使用している。2.9.1 で修正された 2.9.0 の重大なエラーは、BooleanQuery でスコア計算に BooleanScorer が使われたときに、ヒットすべきドキュメントがヒットしないというものである。このエラーは 2.9.0 がリリースされるとまもなく複数のユーザから不具合の現象が報告され、すぐに修正された。このあたりのスピード感はユーザ数の多い OSS ならではといえるだろう。

2. Java ベースのレプリケーション

Solr のインデックスレプリケーション機能は、これまで Unix 系 OS の技術を使ったシェルスクリプトで提供されていた。「Unix 系 OS の技術」とはたとえばファイルシステムのハードリンクや rsync などである。これでは Solr の大きな特徴とされるレプリケーションが Windows では使えず残念である（なにしろ Solr の"r"は Replication の"R"なのだから）。そこでレプリケーション／バックアップも Java で実装してしまおうという提案がなされ、Solr 1.4 で提供されることとなった。Java で書かれているため Solr 本体とシームレスにつながっており、管理画面でモニタリングが可能であり、また各種設定ファイルまでもがレプリケーション可能となっている。

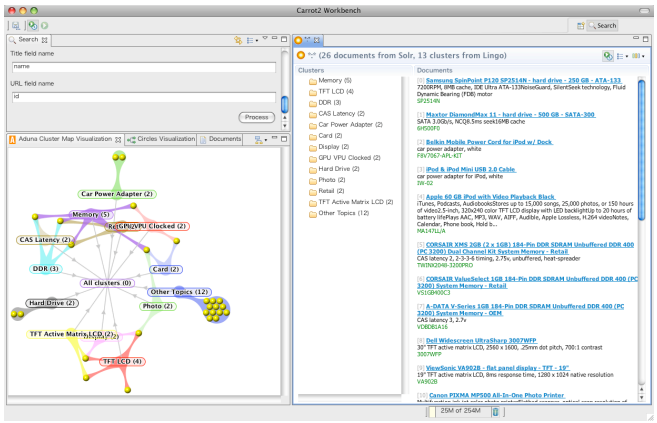
3. 高速ファセット

多値またはトークナイズされ複数単語がインデックスされているフィールドのファセットは、これまで Lucene の FieldCache が使えないために TermEnum を使ってファセット処理をしており時間がかかっていた。そこで docId から各ドキュメントの単語一覧を引くための新しい構造 UnInvertedField が Yonik Seeley 氏によって発明された。UnInvertedField を使ってすべて処理をメモリ上で行うことで高速に処理が可能となった。

4. 新しい SearchComponent

Solr 1.4 でいくつか新しい SearchComponent が追加になった。TermsComponent は検索語のサジェスチョンなどに使うことができる機能を提供する。TermVectorComponent はドキュメント毎の TermVector 情報を返せるようにしたものである。StatsComponent は検索結果の特定数値フィールドの合計や平均を取得できる。たとえば賃貸物件を検索できるサイトで使う場面を考えてみよう。StatsComponent を組み合わせて使うと、「東京都

江東区」「50 平米」「築 3 年以内」という絞り込み検索結果を取得すると同時に、物件価格の相場情報などをユーザに提示できる。ClusteringComponent は検索結果のクラスタリングを可能にする。下図は Solr 1.4 のクラスタリングで使用されている Carrot2 のツールで Solr 付属のサンプルデータのクラスタリングを行った様子を示している。



5. Word/Excel/PDF のサポート

1.3 までの Solr は、インデクシング対象のドキュメントはプレーンテキストのファイルでなければならなかった（プレーンテキストから Solr の XML ファイルなどを作成する）。そのため、Word や Excel などのバイナリファイルは直接インデックスに登録することはできない。登録するには Solr のクライアント側であらかじめテキストデータを抽出し、Solr のインタフェースに沿った形式（XML ファイルなど）で登録しなければならなかった。

Solr 1.4 からは Apache Tika のライブラリを標準で取り込み、Microsoft Office のファイルや OpenOffice のファイル、PDF ファイルなどを直接 Solr に登録することができるようになった。

6. ロールバックのサポート

インデックスへのドキュメントの追加／更新／削除の未コミットの操作を取り消すためのロールバックが追加された。これは Solr 1.3 ユーザの某顧客のために私が書いたパッチをコミュニティにフィードバックしたものである。

ロールバックがない Solr 1.3 までは、一度ドキュメントを Solr に対して追加／更新／削除してしまうと、それは取り消すことができなかった。コミットせずに Solr を再起動しても勝手にコミットされてしまうので、追加／更新／削除する前の状態のインデックスを別途保管しておくなど運用が大変であった。ロールバックの導入により運用設計はだいぶ楽になった。

おわりに

最後に、Solr 1.5 で予定されている新機能のうち、得票数の多いものを紹介しよう。

＜Field Collapsing＞ 検索結果表示にて特定フィールドの重複部分を折りたたむ機能。
＜LocalSolr＞ 地理空間検索機能の取り込み。
＜階層ファセット＞ 階層構造を持つフィールド（例：商品カテゴリなど）のファセット機能。
＜FastVectorHighlighter＞ Lucene の同機能の取り込み。
＜Solr+Hadoop＞ 大規模分散インデックス作成と分散検索。

（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 1-1-3 AIG ビル B1F
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/