Lucene 2.9 の新機能

2009 年 10 月 23 日

株式会社 ロンウイット
コンサルタント
アッタチョー トウンポン

はじめに

2009 年 9 月 26 日に最新版の「Apache Lucene 2.9」がリリースされた。本書はパフォーマンス向上等を目的として追加された一部の新機能を紹介する。

Apache Lucene について

Lucene は、Java で書かれた高機能なオープンソースのテキスト全文検索エンジン・ライブラリである。検索対象になるファイルを解析し、「インデックス」を作成する。インデックスを使用することによって検索処理を高速化する。

現在の Lucene 2.9 はバーション 3.0 がリリースされる前の最後のマイナーリリースである。バーション 3.0 では Java 1.4 から Java 1.5 へ移行し、非推奨の機能を削除することを予定している。

残念ながら Lucene 2.9 は 2.4.1 からの非互換な部分がある。詳しくは、CHANGES.txt の「Change in backwards compatibility policy」で確認できる。過去の Lucene アプリケーション資産がある場合は、ドロップインを試みるより、Lucene 2.9 がある状態で再コンパイルする方がいいだろう。

Quick Background

これから紹介する Lucene 2.9 の新機能について、まず簡単な知識を説明しよう。

Lucene のインデックスデータベースは、たくさんの区分された"セグメント（Segment）"からなる。これらのセグメントには Terms、ポジション、ストアされたテキスト等の情報がそれぞれのファイルに格納されている。ドキュメントをインデックスに追加するとき、Lucene は新しいセグメントを作成し、必要に応じてマージする。

ソートに使用される FieldCache は、フィールド値をキャッシュする為の構造である。例えば、int、long、String 等の基本データ型を格納・解釈することができる。

Lucene のほとんどのアプリケーションでは、インデックスの更新時においてセグメントの変更がない。つまり、多くのメモリ上の内部構造の変更がないのである。 そのため、インデックスの更新後の検索では、インデックス情報のメモリへのリロードを極力少なくすることを目指すことで GC を抑止し、総合的なパフォーマンス向上に寄与することができるのである。

New Features

Lucene 2.9 の新しい機能を以下に紹介する。

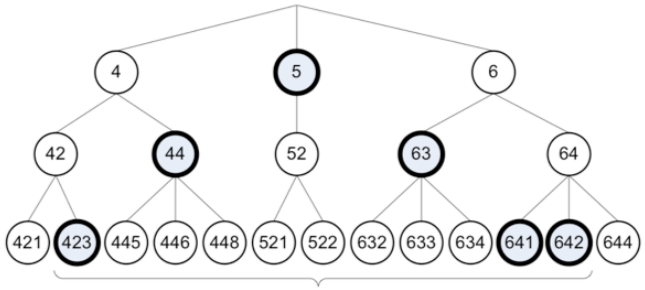
● 高速な IndexReader の取得

Lucene 2.9 では IndexWriter.getReader()というメソッドが導入された。これによりほぼリアルタイムにインデックス更新後の検索ができるようになった。IndexReader.reopen()も併用することができる。さらに IndexWriter.setMergedSegmentWarmer() を使用することによって、セグメントを"暖めて"おくことも可能である。

● 数値の範囲検索

Lucene 2.9 の NumericRangeQuery により数値フィールドの範囲検索が高速化された。

これまで（Lucene 2.4 以前）は、フィールド値の種類が多いフィールドを範囲検索すると、ディスク I/O が増え、検索時間がかかってしまうという問題があった。Lucene 2.9 ではインデックス上にトライ木構造を展開し、範囲検索のスピードを向上させることができた。



● 新しい QueryParser

contrib/queryparser ディレクトリの下に新しくより柔軟な QueryParser フレームワークが追加された。これまでの QueryParser はクエリの字句解析・構文解析を行うための部分が、JavaCC で書かれたソースから生成された Java プログラムであったため、拡張できる部分に制限があった。新しい QueryParser フレームワークでは、カスタマイズ可能な拡張プログラマチック・コントロールを提供し、QueryNode と呼ばれる中間の表現と構文(syntax)を分離している。

なお、この新しい QueryParser は、3.0 で Lucene Core に採用される予定である。

● 新しい Attribute ベースの TokenStream API

Tokenizer と TokenFilter のベースクラスである TokenStream が新しくなった。インデクシング時や検索時には、テキストを分析し、Token に変形する必要がある。Tokenizer は、Reader からもらったテキストを Token に分割し TokenFilter に渡す。
TokenFilter は、Tokenizer が作った Token を正規化したり加工したり、といったことを行う。

この過程で生み出される Token はこれまでの Lucene では固定の属性を保持するクラスであった。すなわち、Term テキスト、オフセット情報、位置増分、ペイロードなどである。2.9 では TokenStream が AttributeSource のサブクラスとなった。
AttributeSource は拡張可能な属性をサポートするが実行時のキャストが不要でそのため性能的にも優れている。また、Token の再利用を促進している点でも GC を押さえることに貢献しており、やはりパフォーマンス向上に一役買っている。また、インデックスに登録するデータをカスタマイズできるようにする将来構想「フレキシブルインデクシング（LUCENE-1458)」をも視野に入れた機能となっている。

● Collector

Lucene 2.9では、HitCollectorを非推奨にし、代わりにCollectorを使うようになった。

Lucene 2.9のCollectorでは、collect()メソッドとscore()メソッドが分別された為、collect()が呼ばれるときのスコア計算をスキップすることができるようになり、スコアが不要な場面では検索時間を高速化できる。

CollectorはドキュメントIDを引数にするcollect()メソッドを持ち、Luceneがヒットしたドキュメントを見つけると呼ばれる。Collectorを実装したアプリケーションは、必要時応じてScorerのscore()メソッドを呼び出すことができる。

●地理空間情報

地理空間情報は、Geo-Spatial(Geography-Spatial)という機能である。この機能は、地理や地図を検索する為に使われている。つまり、距離に基づいた情報や場所等を簡単に検索することができる機能である。たとえば、家から 10km 以内にあるレストランを検索したいとき、Geo-Spatial は緯度経度情報を利用

して位置が条件にマッチするレストランを表示できる。

緯度/経度の Geo コードシステムとして Geohash（http://en.wikipedia.org/wiki/Geohash）が使われており、インデックスサイズ削減の為文字列を後ろから徐々に取り除くことが可能である。

● より使いやすくなった日本語サポート

最後に、弊社関口が「非 ASCII 言語圏のユーザにとって Lucene がより使いやすくなる」ための機能を Lucene 2.9 に提供したので簡単に紹介しよう。

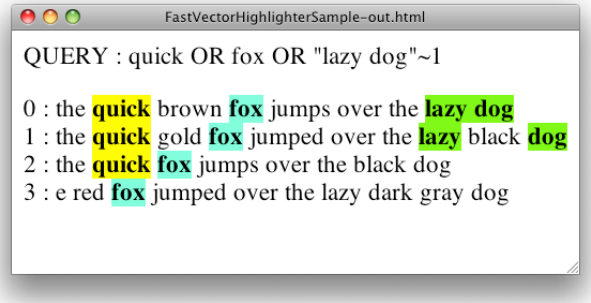
CharFilter

これまでの Lucene は Reader からの文字列データを Tokenizer がトークナイズし、その後、複数の TokenFilter が Token を変形したり切り捨てる、といったフィルタリングを行っていた。しかしこの方法では Token 単位のフィルタリングに限定されてしまうため、文字単位のフィルタリング（例：半角カタカナを全角カタカナに正規化する）を行うためであっても TokenFilter として実装する必要があった。

新しい CharFilter フレームワークは、Reader を Tokenizer に渡す前に文字単位のフィルタリングを行うことを可能にした。さらに CharFilter ではオリジナルのオフセット情報を記憶してオフセット自動補正を行える仕組みを内蔵している。これにより、"ﬆ"（1 文字）を"st"（2 文字）にマッピング（正規化）してもハイライトがずれないようになっている。

FastVectorHighlighter

当初は contrib/highlighter2（略称：H2）として提案されていたが、コミット間際になって Mark Miller 氏によって FastVectorHighlighter と名付けられた。名前の通り、Lucene の TermVector 機能を利用して巨大な文書であっても高速にスニペットを返せるハイライタである。また CJK のような単語間がスペースで分かち書きされない言語圏で多用される N-gram の Analyzer を使用している場合でも正確なハイライトが可能である。また、図のような「多色タグ」でのハイライトが標準で装備されている。さらにスロップを考慮したフレーズ単位のタグ付けも行っている。



まとめ

以上紹介した Lucene 2.9 の新機能は一部であり、すべての変更内容については、CHANGES.txt を参照していただきたい。

Lucene 2.9 は、以前のバーションより多くの新しい機能が追加され、よりいっそうの高性能化と多くの不具合修正がほどこされている。

参考文献

- http://www.lucidimagination.com/
- Lucene in Action, 2nd Edition
- http://lucene.jugem.jp/

（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/