RONDHUIT

Solr サブスクリプション・パッケージ

Solr サブスクリプション・パッケージ（以下パッケージ）は、最新の安定稼働版 Apache Solr を企業ユーザの皆様に、よりいっそう安心・簡単にご利用いただけるように、ロンウイットが独自にパッケージにしてご提供しているサービスです。Solr のサポートサービスはもちろんのこと、ロンウイットのコンサルタントが業務を通じて得た最新の知識や経験が実際のプラグイン形式で提供されるため、お客様はビジネスロジックの開発に集中できます。

パッケージには、以下のものが含まれます。

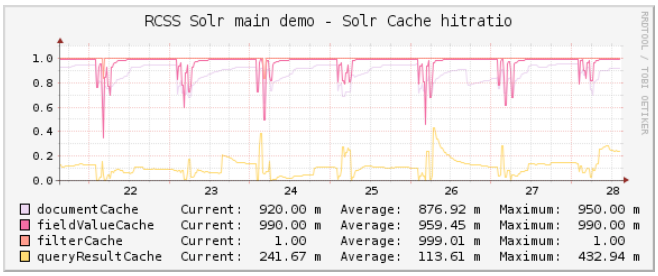
- Tomcat にデプロイずみの安定稼働版 Solr
- Solr サーバ稼働監視
- サンプル環境設定とサンプルデータ
- 日本語に対応したもしかして検索やサジェスチョン、リアルタイムクラスタリング
- 日本語処理プラグイン
- 専門用語抽出／類義語辞書自動作成／形態素解析器／コーパス学習による絞り込み軸検出 などの自然言語処理ツール “NLP4L”
- HTML ノイズ削減ツール
- パーソナライズ検索
- 日本語マニュアルとインストールスクリプト
- セキュリティ情報を含む各種アナウンスとメンバー専用技術情報ページを閲覧するためのアカウント
- サポートサービス

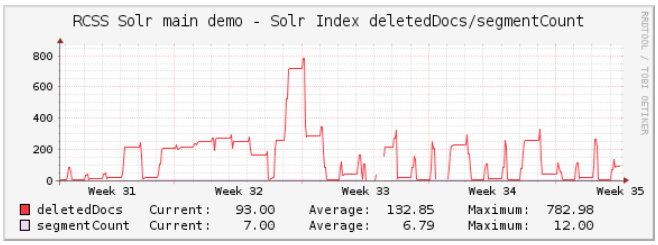
簡単なセットアップ

パッケージは日本語対応を施した Tomcat にデプロイずみの安定稼働版 Solr を含んでいるので、インストールスクリプトを実行するだけで簡単に使い始めることができます。また、さまざまな利用状況に応じたサンプル環境とサンプルデータが同梱されますので、付属の日本語マニュアルを読みながら最新の検索技術に触れることができます。

Solr サーバ稼働監視

CPU/メモリ/Disk/ネットワークなどの OS レイヤだけでなく、JVM ヒープや Solr のキャッシュヒット率やリクエストハンドラ／アップデートハンドラ／レプリケーションハンドラの利用状況まで統合的にモニタリングできます。以下はモニタリング可能な項目の一部です。





105-0003 東京都港区西新橋 1-18-6 クロスオフィス内幸町 11 階
TEL:03-5288-5927
FAX:03-5288-5928

きめ細かな日本語処理プラグイン

日本語の文章には独特の「表記揺れ」があり、これらをきちんと処理しないと、「検索漏れ」が発生してしまいます。「検索漏れ」はお客様の業務によっては、深刻な事態を招いてしまうことがあります。たとえば、EC サイトなどの Web サイト内検索では「検索漏れ」の発生が潜在顧客を逃してしまうことになり、将来利益を逸失することにつながります。

ロンウイットの日本語処理プラグインは、以下のような日本語独特の表記揺れを吸収し、検索漏れの事故を未然に防止します。

半角／全角	ｱｲｳ１２３⇔アイウ123
新旧漢字	慶應大学⇔慶応大学
踊り字	時々⇔時時、部分々々⇔部分部分、いすゞ自動車⇔いすず自動車
漢数字と算用数字	四七／四十七／四拾七⇔47
和暦と西暦	昭和六十四年／平成元年⇔1989 年
読み	かたかな⇔カタカナ⇔katakana、 日本語⇔にほんご
外来複合語	オープンソース・ソフトウェア⇔オープンソースソフトウェア
地名／漢字送りが旭が丘⇔旭丘／卸売り⇔卸売／な	下請け⇔下請／垂れ幕⇔垂幕

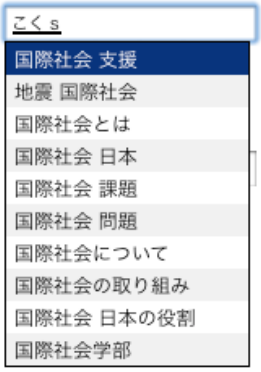
日本語対応「もしかして検索」

キーボードのタイプミスやかな漢字変換のし損じによる検索語エラーに対し、近い検索語を推定して「もしかして○○」のようにユーザに提示できます。リンクとともに正解キーワードが表示され、ユーザはリンクをクリックするだけで再検索できるので、省力化に役立ちます。

貴社業務の専門用語を抽出し徹底活用

Word や PDF など、「自然文」を多く含むドキュメントを検索するシステムを構築する場合、ドキュメントのメタデータが不足しているために、Solr の特徴であるファセット／クラスタリング／類似文書検索などが困難になります。

パッケージに含まれる専門用語抽出機能を用いて文章から「特徴語」すなわち「メタデータ」を引き出してドキュメントにフィードバックし、ドキュメント本来の情報能力をフル活用しましょう。これにより、ファセット／クラスタリング／類似文書検索機能やサジェスチョン（下図）を十分に活用できるようになります。



パーソナライズ検索

Solr は他の検索エンジン同様、ユーザクエリとドキュメントの類似度を計算し、そのスコアの高い順に順位付け（ランキング）をして検索結果を表示します。しかしこの方法では EC サイトで 20 代女性と 50 代男性が「バッグ／鞄」を検索したときに、同じ順番で商品が提示されてしまいます。

パーソナライズ検索は同じユーザクエリでも検索する人によってランキングをより適切なものに調整することができる機能です。これによりユーザは求めている商品（ドキュメント）を素早く探すことができます。

NLP4L

NLP4L は Lucene（Solr も含む）のために開発された自然言語処理ツールです（NLP4L＝Natural Language Processing for Lucene）。NLP4L はお客様がすでに Solr のインデックスに保有している自然言語データの能力を引き出し、活用するのを助けます。Solr インデックスをコーパスとして活用するため、追加工数を最小限に抑えられます。辞書型コーパスから Solr で使える類義語辞書を自動生成したり、再現率と精度の

向上を目指した N-gram 不要で ipadic／Juman／Unidic がすべて利用できる形態素解析器、コーパス学習により絞り込み軸を検出し非構造化文書を構造化文書に転換する機能が提供されます。

Solr は SynonymFilter を使って類義語検索ができます。しかし、類義語辞書は手作業でテキストファイルを編集する必要があるため、継続的なメンテナンスが困難です。そのため、運用時間の経過とともに再現率は落ちていきます。NLP4L の類義語辞書の自動生成機能を使えば、手作業でのメンテナンスを最小限に抑えることができます。類義語辞書は Wikipedia などの汎用的な辞書型コーパスから自動生成できるほか、お客様が管理している商品データやユーザコメントなどからより業務ドメインに特化した類義語辞書を生成することもできます。適当な辞書型コーパスをお持ちでない場合は、Solr のインデックスから専門用語や重要語を抽出し、それらのキーワードを使って Wikipedia などの汎用データベースの部分集合をまず切り出します。そしてその部分集合から類義語辞書を作成すれば、お客様の業務ドメインに近い精度の高い類義語辞書を作成できるでしょう（別途コンサルティングサービスでのご支援をさせていただく必要がございます）。

NLP4L はまた、検索 F 値の向上を目的とした形態素解析器 JaNBestTokenizer もご提供します。F 値は検索システムの性能指標である精度と再現率の調和平均です。通常、JapaneseTokenizer などの形態素解析器を使って Solr を運用しているとき、再現率を上げるために文字 N-gram を併用します。しかし文字 N-gram の併用は精度の低下も招きます。このようなとき、日本語の単語分割の曖昧性に対応した JaNBestTokenizer を使うと、精度の低下を最小限に抑えつつ再現率を向上させることができます。たとえば、「ここではきものを脱ぐ」という文章は、「ここ/では/きもの/を/脱ぐ」と「ここ/で/はきもの/を/脱ぐ」という単語分割の可能性があります。JaNBestTokenizer は、単語分割の上位 N 個のパス（N-best パス）を認識し、第 1 位のパス上のすべてのトークンと、2 位以下のパス上のすべてのユニークな名詞トークンを出力します。これにより、「きもの」も「はきもの」も両方検索にヒットするようになり、文字 N-gram を併用せずに再現率を向上させます。

メンバー限定ページへのアクセス

Solr の応用的な使い方、プログラミングテクニックおよび性能データ等をメンバー限定ページにて公開しています。下記はこれまでに公開された記事の一部です。

- Mahout によるクラスタリング処理の AWS での実行コスト
- 日付時刻データの管理
- optimize の功罪
- グローバル企業での必須機能～言語判別（Solr 3.5/4.0）
- Solr 3.3 で英語をもっとかしこく検索する
- schema.xml のバージョン – 知らないと大いにとまどう Solr 3.3 での変更
- 分散検索への移行 – 巨大なインデックスを分割する
- リラックスクエリー – 検索条件を緩和する
- リアルタイム商品在庫検索 – 希少在庫のランクを下げる

パッケージでご提供するソフトウェアのセキュリティ関連情報をタイムリーにメールでお知らせしています。Solr はもちろん、Tomcat で発見された各種脆弱性やその対応方法までも含みます。

サポートサービス

パッケージを用いて検索システムを開発・運用するお客様に対し、技術サポートをご提供します。開発中に遭遇する Lucene/Solr に関するさまざまな疑問点や運用中に発生した不具合についてのお問い合わせをお受けし、ロンウイットの技術者がお客様の問題解決のために的確なアドバイスをさせていただきます。

サービスレベル契約

受付時間	平日 9:00 – 17:00
一次回答	6 営業時間以内（※）
手段	メールおよび電話
インシデント数	年間 1 2 インシデントまで
サポートバージョン	• 最新の安定バージョン • 以前の安定バージョン（※）

（※）詳細については、弊社ホームページをご覧下さい。

（2014.10.16 版）