

# Rosette と Solr アプリケーション

## オープン・ソース・ソフトウェアを使用して、多言語検索アプリケーションを、高投資効率で作成

ウェブ検索エンジンや、社内向け検索エンジンに広く使われているRosette多言語文書解析技術が、オープンソースの Apache Solr、Apache Lucene でも利用可能になりました。

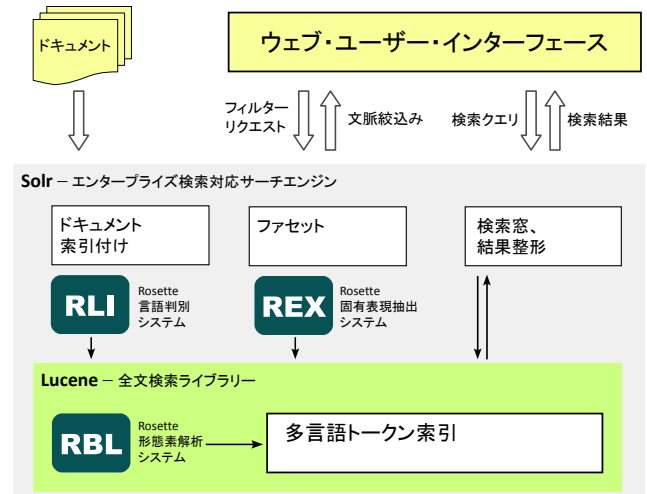
Apache Solr は、一致箇所の強調表示、ファセット型検索、キャッシング、レプリケーション機能、RDBMS インテグレーション、ウェブ管理画面などの特徴を備えた検索サーバーです。Apache Solr は、Java で書かれたオープンソースの情報検索ツールキット Lucene Java の上に構築されていますが、Delphi、Perl、C#、C++、Python、Ruby、PHP 等のクライアント・ライブラリーが用意されており、Java 以外の言語でもアプリケーションを作成できます。Lucene のインデックスはプラットフォーム互換のため、各種ハードウェアやオペレーティングシステムで簡単に使うことができます。Solr と Lucene は、CNET、IBM、Netflix、Wikipedia などの何千もの大規模検索サイトで採用されています。

### 日本語検索

日本語など文章が分かち書きされない文書を検索するには、単語を切り出す言語解析が必要になります。Solr/Lucene に標準で用意されている CJKAnalyzer は、bigram によるもっとも基本的な解析で、高度な検索には必ずしも適しているとは言えません。bigram では単語を二文字ずつ切り出しますから、例えば「東京都」が「東京」でも「京都」でもヒットしてしまいます。提供する Rosette は、本格的な形態素解析ですので、このような問題がなく、精度の高い検索が可能です。

例文	東京都の観光地
bigram	東京 京都 都の の観 観光 光地
形態素	東京 都 の 観光地

形態素 vs. bigram



Rosette は Solr と簡単に融合できるよう設計されており、堅牢で正確な多言語検索技術を、短い工数で実装できます。SDK またはランタイム・パッケージをダウンロードし、いくつかの設定ファイルを編集するだけで、日本語だけでなく、中国語、韓国語、アラビア語、欧州言語など、Rosette の高度な多言語機能をお使いになれます。Solr 側の変更は一切不要で、Rosette が対応する言語の検索が行えるようになります。

### 多言語検索の実装

Rosette は、多言語検索アプリケーションに必要な以下の機能を提供します。

- 言語判別・・・文書の言語と文字コードを自動的に識別します。
- セグメント化(分節)・・・入力文中の単語相当の単位に分割します。
- 基本語化・・・動詞や形容詞の活用形から辞書見出し語の形を取り出します。
- 複合語分割・・・複合語を構成要素に分割して、柔軟な情報検索を実現します。
- 品詞タグ付け・・・各語の品詞を判別します。



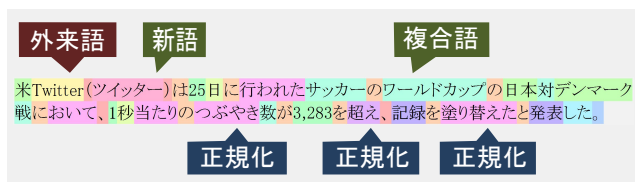
## SOLR の性能と拡張性

以前は高価なメーカー製全文検索エンジンでのみ可能だった、以下のような広範な機能をご利用になれます。

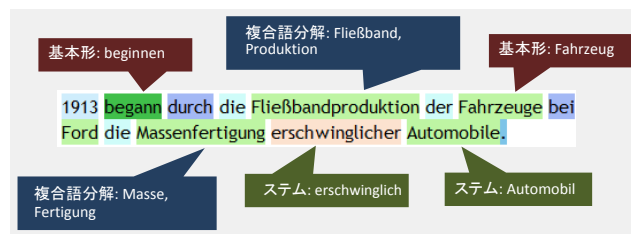
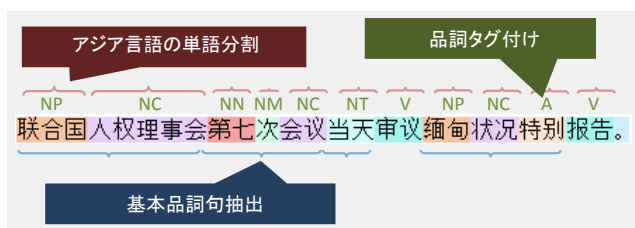
- Windows、Linux、Unix、MacOS をサポート
- 最小限のメモリー
- バッチ処理と同じぐらい高速な増分索引付け
- 元の文書の20%から30%の索引サイズ
- 高度な検索機能と順位付けのアルゴリズム

## その他の多彩な機能

- **日本語正規化**・・・表記ゆれを正規化して、高度な日本語検索を実現できます。
- **名詞句抽出**・・・名詞句を探し出します。
- **停止語削除**・・・「雑音」となる語を取り除くことにより、索引の大きさを小さくします。
- **ユーザー定義辞書**・・・専門語のリストで、標準辞書を補完します。
- **言語領域判定**・・・入力文書中の同一言語領域を特定することにより、適切な言語アナライザーを適用することを可能とします。
- **文節検出**・・・各言語領域内の文節を特定します。
- **中国語文字変換**・・・簡体字と繁体字間の変換により、凡中国語検索を実現できます。文字単位の変換にも語単位の変換にも対応します。



Rosette は、処理言語に応じた最適なアルゴリズムを採用しております。語彙的手法、発見的規則、統計モデルの組み合わせを使用して、速度と精度のバランスを実現しています。



## 処理可能な言語

以下のどの言語の索引付けにも、共通一元化したAPIをご利用になれます

日本語	ギリシャ語	ハンガリー語
朝鮮語	クロアチア語	フィンランド語
中国語(簡体字)	スウェーデン語	フランス語
中国語(繁体字)	スペイン語	ブルガリア語
英語	スロバキア語	ヘブライ語
アラビア語	スロベニア語	ペルシア語
アルバニア語	セルビア語	ポルトガル語
イタリア語	タイ語	ポーランド語
インドネシア語	チェコ語	マレー語
ウクライナ語	デンマーク語	ラトビア語
ウルドゥ語	ドイツ語	ルーマニア語
エストニア語	トルコ語	ロシア語
オランダ語	ノルウェー語	
カタロニア語	バシュトウ語	

## 対応プラットフォーム

以下のプラットフォーム対応のSDKを提供します。その他のプラットフォームのサポートも、ご要望に応じ対応します。

AIX 6.1, PPC	Linux Ubuntu 10.x/11.x, IA32/AMD64
HP-UX 11i, IA64	MacOS
Linux CentOS 4.x/5.x, IA32/AMD64	Solaris 10, SPARC32/64, IA32/AMD64
Linux Debian 5.x, IA32/AMD64	Windows XP/Vista/7, IA32/AMD64
Linux Red Hat 4.x/5.x, IA32/AMD64	Windows Server 2003, 2008

## お問合せ

さらに詳しい製品情報ならびに評価版のご利用をご希望の方は下記へご連絡ください。

[info@basistech.jp](mailto:info@basistech.jp)

[www.basistech.jp](http://www.basistech.jp)

電話 03-3511-2947

詳細 [www.basistech.jp](http://www.basistech.jp)

お問合せ [info@basistech.jp](mailto:info@basistech.jp)

電話 03-3511-2947

〒102-0084  
東京都千代田区二番町9-6

One Alewife Center  
Cambridge, MA 02144

2553 Dulles View Drive  
Herndon, VA 20171

171 Second Street  
San Francisco, CA 94105

