フレームを活用したスクレイピングによるマッシュアップ支援ツール
Web にある情報を構造化して扱いやすく

1. 背景

インターネットが発達して様々な情報を Web サイトに載せるようになった。天気予報や為替相場、株価などの情報を自動的に取得して、別のシステムでデータとして活用したいと望むのは自然な流れだ。ブロードバンドの普及、軽量スクリプト言語の発明、Ajax の整備など、いくつかの条件が重なってマッシュアップというのが盛んになってきている。マッシュアップとは、複数のサイトからデータを組み合わせて新しい価値を生み出すことである。

しかしマッシュアップするためには Web サイトが RSS や Web API を提供していないと不自由だ。そこで Web からデータを取得することができれば、独自の API を作って Web サイトにあるデータを自由に活用できる。やがてブラウザ上で人が目にすることができるものは、コンピューター上のデータとしてすべて利用できるようになるだろう。

2. 目的

これまで企業間の情報のやりとりをするために XML や SOAP といった標準化が進められ、企業内のデータにアクセスするのに SQL や XPATH という手続き言語、ODBC や JDBC のソフトウェア接続方法、接続を抽象化する ADO などが標準化されてきた。

一方で、Web ページに公開されているデータを活用できないかという試みから、スパイダーリングやスクレイピングという技術が生まれた。スパイダーリングとは、Web ページを徘徊して情報を集めてくるソフトウェア技術である。スクレイピングとは、Web ページを自分に都合のよいように情報を加工する技術である。この技術を使うと、例えば、複数のショッピングサイトから商品の値段を調べて比較対象にすることや、地域の不動産会社の情報から集約した不動産情報サイトを構成することが可能である。

本開発の目的は、企業内・企業間のデータのやりとりを別の手段で行う方法を提供する。具体的には Web ページからデータを取り出し、それを構造を持ったデータとして出力をする。その結果マッシュアップされるデータとなり他のシステムで活用する事が可能になる。さらに、データを取り出すだけでなく、Web に対して追加・修正・削除などができる双方向のデータ通信をすることを実現する。

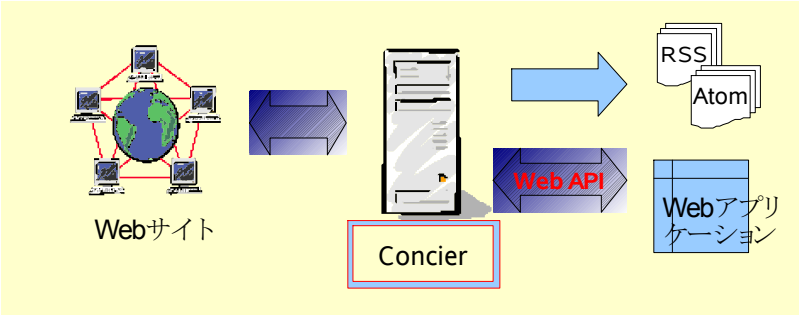

図1　ソフトウェアの位置づけ

3. 開発の内容

本ソフトウェア（Concier）は Ruby 言語および Ruby on Rails をベースに開発しました。

以下の3つの機能を持っています。

1. 記録
   データをとりたい Web ページに本ソフトウェアからアクセスして記録をとります。
2. 編集
   Web ページのどの場所のデータを恒常的に取りたいか設定します。
   検索するキーワードやユーザーID などを入力パラメーターとして設定して、Web ページの特定の場所を出力するように設定します(図1)。
3. 実行
   このサイトに対して、REST でアクセスすると編集したとおり出力されます。出力方法は、CSV, テキスト, RSS などから選ぶことができます。

図 2 「編集」で出力設定をしている画面

そしてプログラム全体の構成は以下の図のように 4 つの部分に分かれます。
当初は1つで構成しておりましたが、分散処理で稼動した方が利点があるので、4つに分けました。分離したことで、それぞれのプログラムを代替することは容易になります。Concier アプリケーション本体も機能に関して以下の A, B, C より成り立っています。

1. Concier アプリケーション本体
   A) Web アクセス部分
   B) 分析・編集部分
   C) 出力部分
2. Web クライアント

3. ベイジアンフィルタ
4. ルールエンジン

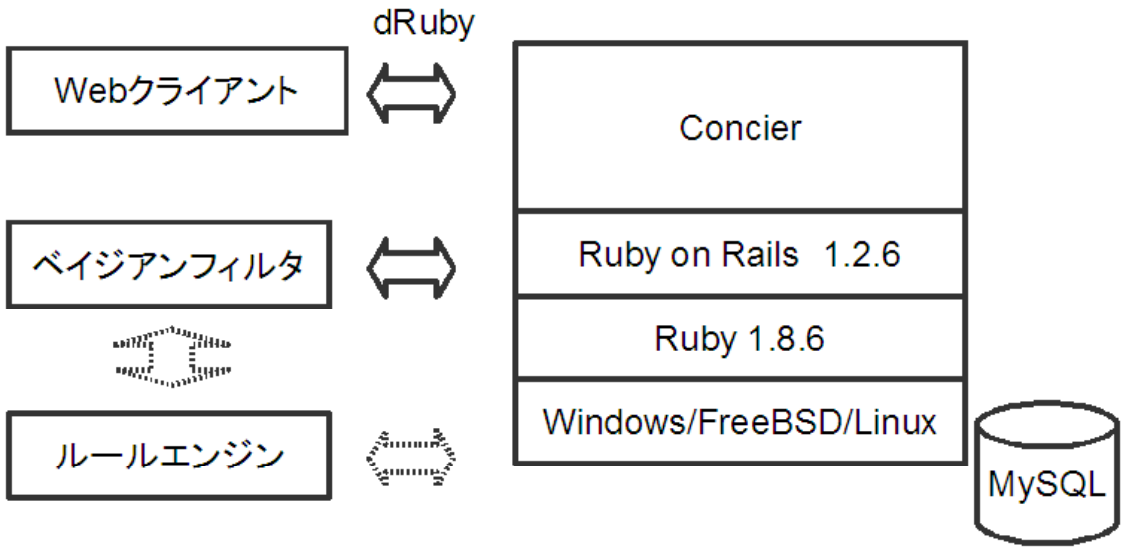

図 3 ソフトウェア構成図

4. 従来の技術（または機能）との相違

- 他のマッシュアップソフト（Plagger, Yahoo Pipes）などと関連性があるが、そのデータの入力を補完するものであって競合するものではない。
- データを送受する方法が異なるが、XMLデータを交換する DataSpiderなどデータ交換するソフトが競合になる。

5. 期待される効果

- インターネットにある世界中のデータをデータベースとして、データ入出力する事が可能となる
- データを統合するマッシュアップの支援
- 異機種間のデータ通信
  データベースやシステムに関係せず、Web ページにデータを入出力できれば、それぞれのシステムにあるデータを双方向で通信する事が可能となる

6. 普及（または活用）の見通し

- インターネットの Web ページは単語による検索が主流となっている。次の世代は Web にある情報を活用していくことが重要となるので、Web ページを構造化して活用していくこのシステムは徐々に受け入れられていくと思う。

7. 開発者名（所属）

大橋 猛（ソリスアート）

（参考）開発者URL
http://www.solisart.com/