セマンティックWeb技術を利用したマッシュアップツールの開発

西村紅美†　塚本享治†
東京工科大学大学院バイオ・情報メディア研究科メディアサイエンス専攻†

1. はじめに

インターネット上の膨大な情報やサービスを組み合わせる方法として、マッシュアップがある。しかし、サービスごとに記述方法や使われている用語が異なっているために、そのままでは組み合わせることが難しい。この問題を、セマンティック Web 技術を用いて解決するマッシュアップツールを開発したので報告する。

2. アプローチ

2.1. マッシュアップにおける問題点

インターネット上にある情報やサービスを組み合わせるためには、以下の問題がある。

(1) 大半の Web ページは HTML である。HTML は仕様に沿っていないものが多く、データ構造が異なる。広告やレイアウトのための情報も多い。

(2) サービスによって使用している用語が異なる。異なるサービスでは、同じ事柄を表していても単語自体は別のものを使用していることが多い。

(3) サービスや情報は頻繁に変更される。そのたびにデバッグが必要である。

2.2. セマンティック Web 技術による解決

これらの問題を解決するために、セマンティック Web 技術を利用して解決を図る。まず、HTML を RDF に変換してデータ構造を統一する。次に、オントロジーを記述し利用することで、用語を統一する。

3. ツールの設計

3.1. ツールの構成

2.節で述べたアプローチを実現するマッシュアップツールを開発する。このツールは、インターネットから情報を集める機能と、OWL による推論検索を行う機能の 2 つのコンポーネントから成り立つ。情報収集コンポーネントは、サービスとの通信と Web ページの RDF 変換を行う。推論検索コンポーネントは、オントロジーによる推論検索を行う。この 2 つのコンポーネントを組み合わせてマッシュアップを行う。デバッグは、コンポーネントの Web ページを利用して行う。[2]

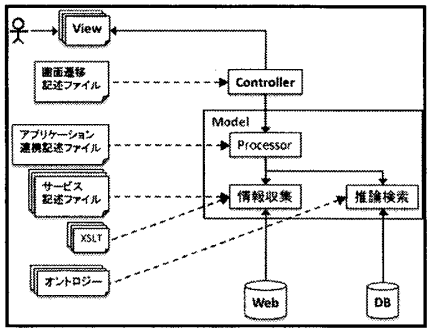

図 1 ツールの構成

3.2. スクリーン・スクレイピング

インターネットの情報の多くが HTML である。HTML から必要な情報のみを抽出する必要がある。しかし、HTML は仕様にそっていないものが多く、広告やレイアウトのための情報が多い。そこで、まず、HTML を XML に変換して仕様の曖昧さをなくす。次に、この XML に対して、XSLT を用いて必要な情報のみを抽出する。その後に RDF に変換する。

3.3. オントロジーによる用語の統一

異なるサービスでは、同じ事柄を表わしていても、Web ページで使用されている用語は異なる場合が多い。Web ページから変換した RDF の用語を統一するために、オントロジーを用いて OWL による推論検索を行う。この時、利用する RDF はいつどこで取得したかが重要である。そこで、RDF に取得日と取得元の URL を付加し、更新の管理に用いる。

3.4. サービスの連携の記述

3.1.節で述べたコンポーネントを組み合わせることで、サービスを組み合わせる。コンポーネントの組み合わせる順序はプログラムに直接記述するのではなく、外部ファイルとして XML で記述する。（図 2）

Development of a mash up tool using Semantic Web Technologies

†Kumi Nishimura, Michiharu Tsukamoto
†Tokyo University of Technology
Graduate School Bionics, Computer and Media Science

```
<?xml version="1.0"? encoding="utf-8">
<config>
 <process>
        <service name="cpu" class="pc-koubou.cpu"/>
        <task>
                <service-from>processor</service-from>
                <case>
                 <from-outcome>sucess</from-outcome>
                 <service-to>pc-koubou.mother</service-to>
                </case>
        </task>
 </process>
 <process>
        <service name="mother" class="pc-koubou.mother"/>
        <task>
                <service-from>cpu</service-from>
                <case>
                 <from-outcome>sucess</from-outcome>
                 <service-to>reasoning</service-to>
                </case>
        </task>
 </process>
 ......
</config>
```

図 2 連携記述ファイル

4. ツールの開発

4.1. サービスからの情報取得

サービスとの通信および Web ページの RDF への変換抽出は既存の情報抽出ツール[1]を利用して情報収集コンポーネントを作成する。このツールは、サービスが持つクッキーやパラメータなどの固有の情報と HTML の変換抽出ルールを XML で記述する。

```
<?xml version="1.0" encoding="utf-8" ?>
<ms:mashup xmlns:ms="http://www.tkzdev.net/2007/mashup">
 <ms:input>
   <ms:param name="searchWord"
             type="text"
             description="検索キーワード"
             value=""/>
   ......
 </ms:input>
 <ms:variable name="search_result">
   <ms:request protocol="GET" request_encoding="shift_jis"
             response_encoding="shift_jis"
             response_form="html"
             url="http://www.pc-koubou.jp/goods/search.php">
     <ms:param name="price" value="$priceRange"/>
     ......
   </ms:request>
 </ms:variable>
 <ms:variable name="url_result">
   <ms:transform source="$search_result" xslt="xsl/cpuList.xsl" />
 </ms:variable>
 <ms:variable name="result">
   <ms:transform source="$url_result"  xslt="xsl/setINTEL.xsl" />
 </ms:variable>
</ms:mashup>
```

図 3 サービス記述ファイル

4.2. OWL による推論検索

作成したオントロジーを用いて、Jena と Pellet による推論検索を行う推論検索コンポーネントを作成する。この時、利用する RDF を RDB に保存して更新の管理を行う。

まず、スクリーン・スクレイピングにより HTML から変換した RDF を RDB に保存する。次に、更新管理のために、この RDF に対して取得日と取得元の URL を記述した RDF を作成し、RDB に保存する。この RDF を利用して推論を行ったのち、推論結果の RDF を RDB に保存する。これを SPARQL 用いて検索することにより利用する。

2 回目以降は、更新情報を用いて前回取得した RDF と比較して、差分があった場合は差分と前回の推論結果とで推論し、RDB に保存する。差分がなかった場合は、前回の推論結果をそのまま利用する。

5. 実験と考察

5.1. 実験

このツールの有効性を検証するために、実際にサービスを組み合わせたシステムを構築した。このシステムは、パソコン工房からパソコンのパーツの情報を利用して、パソコンを自作するときのパーツの組み合わせを提案するものである。[3]構築手順は以下のとおりである。

(1) パソコン工房から情報を集める情報収集コンポーネントを作成する。サービス記述ファイルと Web ページを RDF に変換するための XSLT を作成する。

(2) 推論検索コンポーネントを作成する。パーツの規格のオントロジーを作成する。

5.2. 考察

構築したシステムでは、サービスが提供している情報のデータ構造や用語の違いを気にすることなくサービスを組み合わせることができた。これは、分野を限って実験を行った結果である。パソコンのパーツは規格が決まっており、規格同士の関係がはっきりしているためである。推論結果は、扱う情報がどの事実にそったものなのかによってどの程度信頼できるのか左右される。そのため、RDF を RDB に保存して扱おうとすると、RDB に保存する RDF そのものあるいは複数の RDF を繋げた時に矛盾がある可能性がある。

6. おわりに

分野を限った場合において、セマンティック Web 技術を利用してサービスの組み合わせを簡単に行うことができた。しかし、問題もある。RDB は扱うデータが固定的に定義されている。一方で、インターネットの情報は正確であるかどうかの保証がなく、推論結果に矛盾が生じてしまう可能性がある。

参考文献

[1]山本, 小山, 杉田, 塚本マッシュアップツールの開発とそれを用いた東京都防災マップの構築, 第 69 回全国大会論文, 2007

[2]西村, 塚本, アプリケーション連携システムのスクリーン・スクレイピングを用いたデバッグシステム, SE-163, Vol. 2009, No. 10, pp. 73-pp. 80, 2009

[3]長谷川, 西村, 塚本, セマンティック Web 技術を用いた PC パーツの検索, 第 74 回デジタルドキュメント研究発表会, 2009