6P-6

XML と自然言語解析を用いたレシピポータルサイトの構築

谷村 奈緒子† 山本 剛士‡ 松本 大貴* 塚本 享治†
東京工科大学† 東京工科大学大学院‡ 管理工学研究所*

１．はじめに

インターネット上では、同じ内容の情報であっても、提供するサイトが異なれば表現形式も異なる。例えば、テレビの料理番組では料理番組ごとに様々な工夫がされており、レシピの表現形式が異なっている。しかし、レシピ情報をインターネットで検索するユーザにとって重要なのは Web ページの表現形式ではなく、「何の材料を使い、どのくらい栄養があるのか」といった料理自体の情報である。

本研究では、複数のサイトから提供される料理レシピを対象として、それらのレシピを統一形式に変換し、加工して、あたかも 1 つのサイトのコンテンツのように提供することのできるポータルサイト構築を行った。

２．構築するポータルサイトの全様

システムは、Web 情報抽出スクリプト、栄養価情報の付加、レシピデータベースの 3 つの部分から構成される。システムの全体構成を図１に示す。

まず、非常に複雑な構造になっているレシピ Web ページを各サイトから収集し、Web 情報抽出スクリプト[1]を用いて XML 文書に変換した。取得するサイトは、テレビの料理番組の Web サイトとした。サイトによっては「食材」と「材料」のように同じ意味を表す用語が異なった表現になっているが、この段階で用語と形式を統一した。この XML 文書を「拡張 RecipeXML」と呼ぶ。次に、Web 抽出スクリプトでは解析できない「材料」の要素を多機能日本語処理ライブラリ Ko-BaKo/J[3]を用いて解析し、その結果を使って栄養価の計算を行い、拡張 RecipeXML に埋め込んだ。最後に、データベースに格納してレシピ情報を検索可能にした。

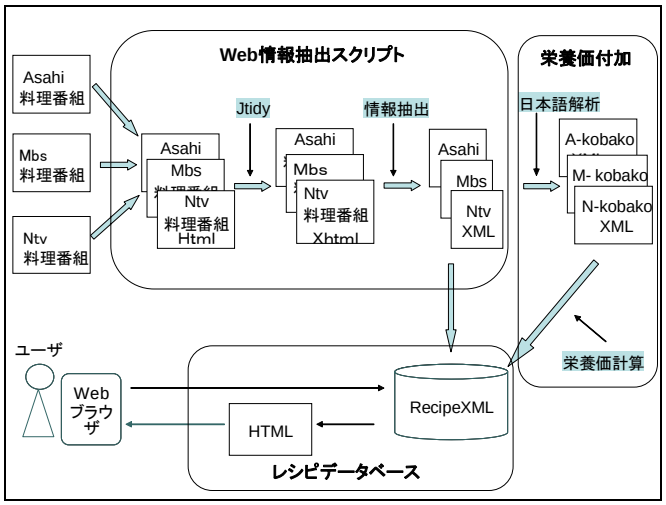

図 1　全体構成図

３．Recipe Markup Language

レシピ情報を取り扱うにあたってインターネット上で公開されている RecipeML[2]を拡張した。この RecipeML は、「料理名」「材料」「分量」「数量」「作り方(ステップ)」の 5 つの要素しか持っていない。これでは、料理番組のサイトが提供する多様な情報を表現することができない。そこで、不足している項目を付け加えて拡張した。追加した項目は、「栄養価情報」や「完成写真」といった、どの料理番組のサイトにも載っている一般的なものである。

４．レシピ Web 情報の取得

レシピ情報は、Web 情報抽出スクリプトを用いて XML 文書に変換した。Web 情報抽出スクリプトは Ant のオリジナルタスクと拡張タスクを用いて、Web サイトから HTML 文書を取得し、XHTML に整形し、さらに情報抽出を行い XML 文書へ変換する(図 1)。

取得した HTML 文書は、図 2 の料理番組「キューピー3 分クッキング」[4]のようにレイアウト情報やコメントアウト情報が大部分を占めており、レシピ自体の情報は 3 割程度しかない。しかし、Web 情報抽出スクリプトを用いた結果、XML 文書は図 3 のように必要なレシピ情報のみを抽出した XML 文書へと変換することができた。

Recipe Portalsite with XML Technologies and Japanese Parser
† Naoko Tanimura , Michiahru Tukamoto – Tokyo University of Technology
‡ Takeshi Yamamoto – Graduate School of “Tokyo University of Technology”
＊ Taiki Matsumoto – Kanrikogaku Kenkyusyo, Ltd.

```
<HTML>
<HEAD>
<!--にほんテレビ-->
<TITLE>1/4　花巻きと中国茶</TITLE>
<META                              HTTP-EQUIV="Content-Type"
CONTENT="text/html;charset=shift_jis">
<META HTTP-EQUIV="Pragma" CONTENT="no-cache">
<META                    HTTP-EQUIV="Content-Script-Type"
content="text/javascript">
<SCRIPT SRC="../font.js"></SCRIPT>
<style type="text/css"><!--
body { margin-top: 10px;
 margin-left: 0px;
 margin-right: 0px;
 margin-bottom: 0px;
```

図 2 HTML(キューピー3 分クッキング)

```
<?xml version="1.0" encoding="Shift_JIS"?>
<レシピ>
<品名>1/4　花巻きと中国茶</品名>
<概要>新年好！お正月のティータイムにおすすめ 花巻きと中
国茶</概要>
<完成写真>images/010400.jpg</完成写真>
<栄養価>
<エネルギー>151kcal</エネルギー>
<たんぱく質>15g</たんぱく質>
<塩分>1.8g</塩分>
・・・
</栄養価>
<食材と分量>
```

図 3 RecipeXML(キューピー3 分クッキング)

５．栄養価情報の付加

栄養価情報は全てのレシピ Web ページに載っているわけではない。しかし健康が重要視される昨今、誰もが気になる情報である。そこで材料とその分量から栄養価を計算し、付加することにした。栄養価情報の付加方法を図 4 に示す。

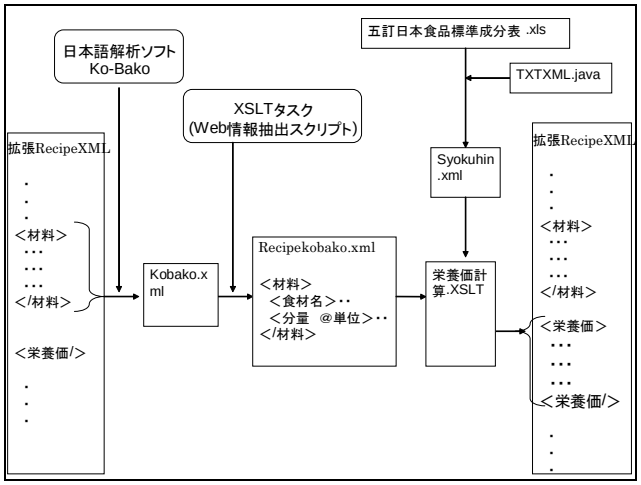

図 4　栄養価情報の付加

まず、栄養価計算の基準となる五訂日本食品標準成分表をインターネット上の栄養価計算ソフトヘルシーメーカー[5]から取得し Excel ファイルだったものを XML 文書に変換し、syokuhin.xml として保管した。

次に、拡張 RecipeXML の材料を syokuhin.xml から探さなければならないが、そのためには現段階の拡張 RecipeXML では「<材料>砂糖 大さじ1<材料>」などのように、材料名と分量が同じタグ内に記されているものを分けなければならない。そこで、多機能日本語処理ライブラリ Ko-BaKo を使用して、形態素解析を行い、単語を区切ることにした。

最後に、栄養価情報を計算する。五訂日本食品標準成分表は食品 100g 中の栄養価情報が記されているため、それを分量に合わせて計算しなければならない。これも、XSL を使って行った。

６．レシピデータベースの構築

拡張 RecipeXML をデータベースに保管し、検索可能にした。データベースは、RDB や xmlDB などもあるが、本研究ではデバックしやすい理由からディレクトリに XML 形式のファイルとして保管する方法を用いた。これは、XML ネイティブデータベースとほぼ同じインターフェースとなっている。

７．おわりに

統一形式のポータルサイトの構築にあたってレシピ情報を取り上げ、インターネット上の HTML 文書を XML 化し、付加情報を付けて、検索可能なシステムを構築した。本システムは各段階において手動で行わなければならないことや、五訂日本食品標準成分表に情報がない材料の対応をいかにするのかなど、課題が多く残っている。今後、多くのレシピ Web サイトを対象に実験を行い、変換率を向上させる必要がある。

なお、各サイトから入手したレシピ情報には著作権があるため、研究のためだけに用い、公開はしていない。

参考文献

[1]松本，塚本，“XSLT スクリプトの対話的な生成”，情報処理学会研究報告 2005-DD-48，2005
[2] FormatData，The Recipe Markup Language，http://www.formatdata.com/recipeml/，2001
[3]㈱日本システムアプリケーション，多機能日本語処理ライブラリ Ko-BaKo/J，2005
[4] NTV，キューピー3 分クッキング http://www.ntv.co.jp/3min/
[5] ㈲ マッシュルームソフト，栄養価計算 ヘルシーメーカー431，2005