テキスト処理における WebAPI の利用

Utilization of WebAPI in Text Processing

高 尾 哲 康
Takao Tetsuyasu

1.　はじめに

近来の Web2.0[1]のめざましい発展により、Web のネットワーク化および構造化が進んできた。これまでの Web アプリケーションでは、サーバで動作する Web や DB アプリケーションなどの利用や Web ブラウザ上で動作する JavaScript や Ajax 等の動的なページを利用してきた。今後は Web 上にあるさまざまなサービスをブラウザからだけでなく、さまざまなプログラムから利用するインタフェースを備えた WebAPI(Application Program Interface)としての利用が広まりつつある。WebAPI を提供している代表的なサイトには、Amazon、Google、Yahoo!、「はてな」などがあるが、さらに先進的な WebAPI を提供するサイトも出現しつつある。これらのサイトでは WebAPI を提供することにより、サービス利用者の増加やそれに伴う顧客層の拡大が大きなメリットとなっている。多くの利用者に認知されつつあるサービスには、Web やブログ、商品情報などの検索サービス、地図情報サービス、ブックマーク共有（ソーシャル・ブックマーク）などがある。いずれも閲覧性の高いブラウザ上でのサービスであり、使い勝手も良く、頻繁に利用されている。しかし、これらの検索結果などを Excel や手元のアプリケーションソフトウェアなどで加工・編集したい場合はブラウザからの単調なコピー&ペーストの作業量が多くなり、せっかくの Web アプリケーションサービスの利用に支障をきたしている。

今回、図書選定リストの作成やゼミ等の演習用問題作成などに各サイトの WebAPI のサービスを活用することにより作業量の大幅な軽減をはかることができた。本報告では、AWS(Amazon Web Service)を利用した図書選定リスト作成、Yahoo!デベロッパーネットワークが提供しているテキスト解析 WebAPI を利用した演習用問題作成やルビ自動付加について紹介する。

2.　Amazon Web Service の利用

教育用図書、新刊図書などの図書選定リストの作成のために AWS(Amazon Web Service)[2]を利用したキーワードによる既刊図書検索プログラムを Perl 言語で作成した。AWS 利用のための署名認証には HMAC-SHA256[3]機能を利用している。検索範囲のジャンル（文学・評論、思想・社会・ノンフィクション、人文・思想、社会・政治をはじめ、合計 30 ジャンル)、検索結果数や並べ替

え方法（売れている順番、価格順、タイトル順、発売日順）などはプログラムの引数として指定する。図１にその実行例を示す。引数として、ジャンルはコンピュータ・インターネット、検索結果数は 10 個で売れている順番で表示するようにしている。ひとつの結果は 1 行表示で、ASIN(Amazon Standard Identification Number)、タイトル、著者名、出版社、ISBN、価格、発行日をタブ区切りなどで出力する。検索結果は XML フォーマットで得られるので、XML テキストを解析して CSV やタブ区切りなどに変換することで Excel などへ組み込むことができる。XML テキストの解析には Perl モジュールを利用することで簡単に処理できる。

```
$ amazon.pl -B11 -S1 Webアプリケーション
479735397X      Photoshop 10年使える逆引き手帖 【CS4/CS3/CS2/CS/7.0対応】 (ああ
したい。こうしたい。)   藤本 圭 ソフトバンククリエイティブ      479735397X      2
520     2009-05-29
4797353961      Illustrator 10年使える逆引き手帖【CS4/CS3/CS2/CS/10/9/8 対応】 (
ああしたい。こうしたい。)       高野 雅弘       ソフトバンククリエイティブ      4
797353961       2520    2009-07-30
4774135666      [24時間365日] サーバ/インフラを支える技術 ?スケーラビリティ、ハ
イパフォーマンス、省力運用 (WEB+DB PRESS plusシリーズ)  安井 真伸,横川 和哉,ひろ
せ まさあき,伊藤 直也,田中 慎司,勝見 祐己       技術評論社      4774135666      2
919     2008-08-07
4797346809      詳解 Objective-C 2.0    荻原 剛志       ソフトバンククリエイティ
ブ      4797346809      4410    2008-05-28
4274067858      RailsによるアジャイルWebアプリケーション開発    Sam Ruby,David H
einemeier Hansson,Dave Thomas   オーム社        4274067858      4410    2009-12
4797348534      Photoshopデザインラボ -プロに学ぶ、一生枯れない永久不滅テクニッ
ク- (Design Lab+ 1-1)   デザインラボ編集部      ソフトバンククリエイティブ      4
797348534       1890    2008-05-30
479733245X      PHPによるWebアプリケーションスーパーサンプル 第2版      西沢 直
木      ソフトバンククリエイティブ      479733245X      3990    2006-03-24
4797356669      Eclipse 3.5 完全攻略    宮本 信二       ソフトバンククリエイティ
ブ      4797356669      2940    2009-10-08
4797354410      GameGraphicsDesign キャラクターCG彩色テクニック 瑞穂 わか
ソフトバンククリエイティブ      4797354410      3129    2009-08-27
4822229866      Webアプリケーション・サーバー 設計・構築ノウハウ      NTTデー
タ先端技術、NTTデータ   日経BP社        4822229866      2940    2008-07-10
$
```

	A	B	C	D	E	F	G
1	ASIN	タイトル	著者名	出版社	ISBN	価格	発行日
2	479735397X	Photoshop10年使える逆引き手帖	藤本 圭	ソフトバンククリエイティブ	479735397X	¥2,520	2009/5/29
3	4797353961	Illustrator10年使える逆引き手帖【(	高野 雅弘	ソフトバンククリエイティブ	4797353961	¥2,520	2009/7/30
4	4774135666	[24時間365日] サーバ/インフラをま	安井 真伸,横川 和哉	技術評論社	4774135666	¥2,919	2008/8/7
5	4797346809	詳解 Objective-C 2.0	荻原 剛志	ソフトバンククリエイティブ	4797346809	¥4,410	2008/5/28
6	4274067858	RailsによるアジャイルWebアプリケー	Sam Ruby,David Hei	オーム社	4274067858	¥4,410	2009/12/1
7	4797348534	Photoshopデザインラボ -プロに学ぶ	デザインラボ編集部	ソフトバンククリエイティブ	4797348534	¥1,890	2008/5/30
8	479733245X	PHPによるWebアプリケーションスー	西沢 直木	ソフトバンククリエイティブ	479733245X	¥3,990	2006/3/24
9	4797356669	Eclipse 3.5 完全攻略	宮本 信二	ソフトバンククリエイティブ	4797356669	¥2,940	2009/10/8
10	4797354410	GameGraphicsDesign キャラクターC	瑞穂 わか	ソフトバンククリエイティブ	4797354410	¥3,129	2009/8/27
11	4822229866	Webアプリケーション・サーバー 設計	NTTデータ先端技術、	日経BP社	4822229866	¥2,940	2008/7/10

図１．AWS を利用した検索と Excel への取り込み

3.　Yahoo!デベロッパーネットワークのテキスト解析の利用

テキスト解析 API[4]には、日本語形態素解析、かな漢字変換、ルビ振り、校正支援、日本語係り受け解析、キーフレーズ抽出がある。今回はルビ付き 2 文文節単位混合文を演習用問題として作成してみた。2 文文節単位混合文とは、2 つの文章をそれぞれ文節単位で区切り、並びの順序を

保ったままランダムにマージした文のことである。問題の作成は次の手順で行なう。

①　問題文となるテキストファイルを複数用意

この例ではWebページなどから新聞のコラム欄のテキストを利用した。コラムごとに1つのファイルとする。問題文の作成にあたっては、各テキストファイルに含まれるテキストの長さにあまり差がない文を選んだほうがよい。

日本経済新聞 2009-07-11 付「春秋」(14 文、534 文字) : 20090711syunjyuu.txt

日本経済新聞 2009-07-13 付「春秋」(17 文、540 文字) : 20090713syunjyuu.txt

②　文節単位に分割（文節/行)。

形態素解析により文を形態素単位に分割し、係り受け解析により文節（自立語＋付属語）単位として構成する。なお、後処理として問題文が難しくならないように複合語、関連性の強い語彙や連体修飾などはひとつの文節にまとめている。

```
$ ydnda.pl -f1 20090711syunjyuu.txt > 20090711syunjyuu_m.txt
$ ydnda.pl -f1 20090713syunjyuu.txt > 20090713syunjyuu_m.txt
$ head 20090711syunjyuu_m.txt 20090713syunjyuu_m.txt
==> 20090711syunjyuu_m.txt <==
目に見えぬ
ロボットが
世界を
荒らし回っている。
韓国や米国の
政府機関や銀行、
証券取引所などの
ウェブサイトを
集中的に狙った
サイバー攻撃のことだ。
==> 20090713syunjyuu_m.txt <==
スポーツジャーナリストの
二宮清純さんによれば、
選手に
一番印象に
残っている言葉を
聞くと、
例外なく
調子の
悪いとき
かけてもらった
$
```

③　ルビ振り

ルビ振りのグレードには「小学校学習指導要領」付録「学年別漢字配当表」(1989 年 3 月 15 日付文部科学省告示。1992 年 4 月施行) を参考に、小学 1 年生～6 年生の各学年向け、中学生向け、および一般向けと 8 段階に設定できる。この例ではグレード 4 (小学 4 年生向け。小学 1～3 年生で習う漢字にはふりがなを付けない) を-g オプションにて指定している。ルビ振り機能は留学生などの外国人向けに利用している。これにより、ルビの部分には HTML の RUBY タグが付けられる。

```
$ ydnruby.pl -g4 20090711syunjyuu_m.txt > 20090711syunjyuu_mr.txt
$ ydnruby.pl -g4 20090713syunjyuu_m.txt > 20090713syunjyuu_mr.txt
$ head 20090711syunjyuu_mr.txt 20090713syunjyuu_mr.txt
==> 20090711syunjyuu_mr.txt <==
目に見えぬ
ロボットが
世界を
<ruby><rb>荒</rb><rp>(</rp><rt>あ</rt><rp>)</rp></ruby>らし<ruby><rb>回</rb><rp>
(</rp><rt>まわ</rt><rp>)</rp></ruby>っている。
<ruby><rb>韓国</rb><rp>(</rp><rt>かんこく</rt><rp>)</rp></ruby>や米国の
<ruby><rb>政府</rb><rp>(</rp><rt>せいふ</rt><rp>)</rp></ruby><ruby><rb>機関</rb>
<rp>(</rp><rt>きかん</rt><rp>)</rp></ruby>や銀行、
<ruby><rb>証券</rb><rp>(</rp><rt>しょうけん</rt><rp>)</rp></ruby>取引所などの
ウェブサイトを
<ruby><rb>集中的</rb><rp>(</rp><rt>しゅうちゅうてき</rt><rp>)</rp></ruby>に<ruby
><rb>狙</rb><rp>(</rp><rt>ねら</rt><rp>)</rp></ruby>った
サイバー<ruby><rb>攻撃</rb><rp>(</rp><rt>こうげき</rt><rp>)</rp></ruby>のことだ。
==> 20090713syunjyuu_mr.txt <==
スポーツジャーナリストの
<ruby><rb> 二 宮 </rb><rp>(</rp><rt> に の み や </rt><rp>)</rp></ruby><ruby><rb> 清 純
</rb><rp>(</rp><rt>せいじゅん</rt><rp>)</rp></ruby>さんによれば、
<ruby><rb>選手</rb><rp>(</rp><rt>せんしゅ</rt><rp>)</rp></ruby>に
一番<ruby><rb>印象</rb><rp>(</rp><rt>いんしょう</rt><rp>)</rp></ruby>に
<ruby><rb>残</rb><rp>(</rp><rt>のこ</rt><rp>)</rp></ruby>っている言葉を
聞くと、
<ruby><rb>例外</rb><rp>(</rp><rt>れいがい</rt><rp>)</rp></ruby>なく
調子の
悪いとき
かけてもらった
$
```

④　MS-Word 読み込み用ルビ付 2 文文節単位混合文作成

この例では 2 文混合文を作成しているが、プログラムでは n 文混合文として作成することもできる。その場合はルビ付きの 1 行 1 文節単位のテキストファイルを n 個用意することになる。-s

オプションには問題文中の文節間の区切り文字を指定する。出力ファイルには MS-Word を HTML 形式で読み込んだ際のスタイルを指定する CSS が埋め込まれる。-r オプションは RUBY タグ付き HTML 化を意味し、-m オプションは形態素解析における形態素や係り受け解析における文節等の区切り済み文字列を対象とする。ルビ文字のスタイルなどもこの CSS にて指定している。生成されたファイルの例を図２に示す。

$ columnmerge.pl -s'・'-mr 20090711syunjyuu_mr.txt 20090713syunjyuu_mr.txt | nkf -Lw -s > 20090714.doc

$

ファイル(F)　編集(E)　表示(V)　端末(T)　タブ(B)　ヘルプ(H)

```
<html>
<head>
<meta http-equiv=Content-Type content="text/html; charset=shift_jis">
<meta name=ProgId content=Word.Document>
<meta name=Generator content="Microsoft Word 11">
<meta name=Originator content="Microsoft Word 11">
<xml>
<w:WordDocument>
<w:DisplayBackgroundShape/>
<w:PunctuationKerning/>
</w:WordDocument>
</xml>
<style>
<!--
/* Style Definitions */
p.MsoNormal
{mso-style-parent:"";
 margin:0mm;
 margin-bottom:.0001pt;
 text-align:justify;
 text-justify:inter-ideograph;
 mso-pagination:none;
 font-size:10.5pt;
 mso-bidi-font-size:12.0pt;
 font-family:"ＭＳ ゴシック";
 mso-bidi-font-family:"ＭＳ ゴシック";
 mso-font-kerning:1.0pt;}
p.MsoPlainText
{margin:0mm;
 margin-bottom:.0001pt;
 text-align:justify;
 text-justify:inter-ideograph;
 mso-pagination:none;
 font-size:12.0pt;
 font-family:"ＭＳ ゴシック";
 mso-font-kerning:1.0pt;
 text-indent:0mm;
 mso-char-indent-count:0;
/* line-height:24.0pt; */
 mso-line-height-rule:exactly;}
rt.MsoRubyText
{font-size:6.0pt;
 layout-grid-mode:line;}
/* Page Definitions */
@page
{mso-page-border-surround-header:no;
 mso-page-border-surround-footer:no;}
@page Section1
{size: 210mm 297mm;
 margin:15mm 15mm 10mm 15mm;
 mso-paper-source:0;
 layout-grid:14.3pt;}
div.Section1
{page:Section1;}
-->
</style>
</head>
<body lang=JA style='tab-interval:42.0pt;text-justify-trim:punct-and-kana;line-b
reak:strict'>
<div class=Section1 style='layout-grid:14.3pt'>
<p class=MsoPlainText>16 - 19</p>
<p class=MsoPlainText>目に見えぬ ・ スポーツジャーナリストの ・ <ruby><rb>二宮</
rb><rp>(</rp><rt class=MsoRubyText>にのみや</rt><rp>)</rp></ruby><ruby><rb>清純<
/rb><rp>(</rp><rt class=MsoRubyText>せいじゅん</rt><rp>)</rp></ruby>さんによれば
、 ・ ロボットが ・ <ruby><rb>選手</rb><rp>(</rp><rt class=MsoRubyText>せんしゅ<
/rt><rp>)</rp></ruby>に ・ 一番<ruby><rb>印象</rb><rp>(</rp><rt class=MsoRubyTex
t>いんしょう</rt><rp>)</rp></ruby>に ・ 世界を ・ <ruby><rb>荒</rb><rp>(</rp><rt
20090714.doc 17%
```

図２．CSS が埋め込まれたルビ付き 2 文文節単位混合文例

なお、区切り済みテキストだけでなく、任意文字数単位で混合文を作成することも可能となっている。こちらのほうが作成された問題文の難易度は高くなる。

⑤　作成したファイル（上記の 20090714.doc）を MS-Word にて読み込み、レイアウトを修正

完成したルビ付 2 文文節単位混合文の例を図３に示す。パラグラフ先頭の「16 - 19」は、2 文を分類したときのそれぞれの文やパラグラフに含まれる文節数を意味する。なお、columnmerge.pl コマンドで出力した RUBY タグ付きテキストを MS-Word 内に直接ペーストしても HTML タグとして認識されないので、ファイル経由で MS-Word から読み込みをする必要がある。また、現在のルビ振り WebAPI では単語単位でのルビ付けであり、文字単位でのルビ付けは行なっていない。

2009 年 7 月 11 日と 7 月 13 日付の日本経済新聞「春 秋」から

16 - 19

スポーツジャーナリストの ・ 目に見えぬ ・ ロボットが ・ 二宮清純さんによれば、 ・ 世界を ・ 荒らし回っている。 ・ 選手に ・ 一番印象に ・ 残っている言葉を ・ 聞くと、 ・ 例外なく ・ 調子の ・ 韓国や米国の ・ 悪いとき ・ かけてもらった ・ 言葉を ・ 政府機関や銀行、 ・ 証券取引所などの ・ 挙げるという。 ・ ウェブサイトを ・ 「どん底時代のことほど ・ 集中的に狙った ・ 記憶に ・ サイバー攻撃のことだ。 ・ 残るんでしょうね」。 ・ 使われているのは ・ 「ボット」という ・ ウイルス。 ・ ロボットを ・ 将棋の ・ 羽生善治さんとの ・ 対談で ・ もじった ・ 言葉だという。 ・ そう語っている。

図３．WebAPI を利用したルビ付き 2 文文節単位混合文作成例

これらの問題文を活用することにより、意味をよく考えながら文章を読む習慣をつけることができると期待される。

4.　まとめ

テキスト処理において WebAPI を利用した方法のいくつかを紹介した。これら以外にも漢字検定や一般常識問題などについて検索や問題集作成アプリケーションを開発した。今後も作業の効率化を図るために WebAPI を活用したさまざまなアプリケーションを開発していく予定である。

参考文献

(1) 小川 浩、後藤 康成、Web2.0 BOOK、インプレス (2006)
(2) Amazon Web サービス http://www.amazon.co.jp/
(3) H. Krawczyk, M. Bellare, R. Canetti、RFC 2104 - HMAC: Keyed-Hashing for Message Authentication　http://www.faqs.org/rfcs/rfc2104.html
(4) Yahoo!デベロッパーネットワーク http://developer.yahoo.co.jp/