サービスコンピューティングのための HTML利用基盤の確立

大阪工業大学 情報科学部
須永　宏

本日の発表内容

●HTMLサイトの利用基盤
（特定の案件ではなくできる限り汎用的に）

●HTML5のグラフ作成機能によるグラフ表示
（サイトより取得したデータからグラフ生成）

HTMLの利用の問題点

●HTMLは，ビジュアル重視でデータ構造など統一規格無し．統計情報サイトなどはデータの宝庫だが，XMLと違い，プログラムでの扱いは困難．

●送信時のFormやパラメタの扱いも判りにくく，パケットモニタでHttpパケットを観測し判断することも．

●JavaScriptが使われているとさらに複雑．

しかし，XML APIで提供されていない有用な情報が無限にある！

XML構造化文書は操作しやすい（DOM）

タグ名指定で該当タグをList形式で取り出し
nodelist = element.getElementsByTagName("pastHistory");

```
<pastHistory>
    <byoumei>ヘルニア</byoumei>
    <timeage>2011年12月</timeage>
    <outcome>良好</outcome>
    <comment>次回再検査</comment>
</pastHistory>
<pastHistory>
    ……
</pastHistory>
```

要素の取り出し
subnodelist.item(i).getFirstChild() . getNodeValue();

XMLの問題点＝APIが少ない！

●一般検索（Yahoo, Bing），翻訳（Google），地図（Google），音楽検索（Youtube RSS），商品検索（楽天）などいくつかの汎用的なサービスは提供されている．

●有料化・会員限定など使いにくくなっている．

●作りたいアプリケーションの目的に沿ったAPIが殆ど無い．

→HTMLサイトから情報を引き出したい！！！

本研究の対象となるHTMLサイトの例

●http://prf.uub.jp/prefbase.html

2012年10月1日現在の自治体数（最新）

都道府県	都道府県庁所在地	推計人口（人）	面積（平方km）	人口密度（人/平方km）	国勢調査人口（人）	市町村数			
						市	区	町	村
北海道	札幌市	5,485,916	78,420.86	69.95	5,506,419	35	10	129	15
青森県	青森市	1,363,006	9,644.55	141.32	1,373,339	10		22	8
岩手県	盛岡市	1,312,756	15,278.89	85.92	1,330,147	13		15	5
宮城県	仙台市	2,323,224	7,285.77	318.87	2,348,165	13	5	21	1
秋田県	秋田市	1,075,055	11,636.28	92.39	1,085,997	13		9	3
山形県	山形市	1,161,294	9,323.46	124.56	1,168,924	13		19	3
福島県	福島市	1,988,995	13,782.76	144.31	2,029,064	13		31	15
茨城県	水戸市	2,956,854	6,095.72	485.07	2,969,770	32		10	2
栃木県	宇都宮市	2,000,021	6,408.28	312.10	2,007,683	14		12	
群馬県	前橋市	2,000,876	6,362.33	314.49	2,008,068	12		15	8
埼玉県	さいたま市	7,204,168	3,798.08	1,896.79	7,194,556	40	10	22	1
千葉県	千葉市	6,211,820	5,156.61	1,204.63	6,216,289	36	6	17	1
東京都	新宿区	13,186,562	2,188.67	6,024.92	13,159,388	26	23	5	8
神奈川県	横浜市	9,059,616	2,415.86	3,750.06	9,048,331	19	28	13	1
新潟県	新潟市	2,362,581	12,583.83	187.75	2,374,450	20	8	6	4
富山県	富山市	1,088,409	4,247.61	256.24	1,093,247	10		4	1
石川県	金沢市	1,166,315	4,185.67	278.64	1,169,788	11		8	

対象HTMLサイトの文書構造

●HTMLソース→<table>タグ

方式提案の着眼点

①<table>構造主体のHTMLサイトには統計情報を扱っているものが多い.

②該当サイトのURLを入力し, HTML文書を取得. <table>をキーにデータを抽出.

③SQLデータベースに格納.

④SQLデータベースのテーブルをXML構造化文書に自動変換, Web APIとして提供.

⑤HTML5によるグラフ生成.

tableタグの取り出し

●URLを投入し, 返却されたHTML文書の中の<table>～</table>までを順次見つける.

<table>～</table>の組を見つけた分だけ順に表示. それ以外は消去.

tableタグブロックの選択

●複数あった<table>～</table>の中から選択されたものを表示する. 列番号・行番号を指定可能なように, 第1行に<tr>～</tr>を挿入し, 各<td>～</td>内にcheck boxを入れる.

最終行に, 行番号を指定するためのcheck boxを加える.

table要素の指定

●check boxで指定された列と行番号の範囲内で, DB投入用のformを作る.

Sqliteライブラリを利用してDB, テーブル生成.

number	都道府県	都道府県・庁所在地	推計人口・(人)	面積・(平方km)
text	text	text	text	text
1	市	区	町	村
2	北海道	札幌市	5,485,916	78,420.86
3	青森県	青森市	1,363,006	9,644.55
4	岩手県	盛岡市	1,312,756	15,278.89

データベース・テーブル生成

●生成されたテーブルは, 各種操作がしやすいように, HTML表示の際には, 列と行にcheck boxを入れている.

Table name = fukendata

number	都道府県	都道府県庁所在地	推計人口(人)	面積(平方km)	人口密度(人・平方km)	国勢調査人口(人)
text	text	text	text	text	text	text
1	2	3	4	5	6	7
1	北海道	札幌市	5,485,916	78,420.86	69.95	5,506,419
2	青森県	青森市	1,363,006	9,644.55	141.32	1,373,339
3	岩手県	盛岡市	1,312,756	15,278.89	85.92	1,330,147
4	宮城県	仙台市	2,323,224	7,285.77	318.87	2,348,165
5	秋田県	秋田市	1,075,055	11,636.28	92.39	1,085,997
6	山形県	山形市	1,161,294	9,323.46	124.56	1,168,924
7	福島県	福島市	1,988,995	13,782.76	144.31	2,029,064
8	茨城県	水戸市	2,956,854	6,095.72	485.07	2,969,770
9	栃木県	宇都宮市	2,000,021	6,408.28	312.10	2,007,683
10	群馬県	前橋市	2,000,876	6,362.33	314.49	2,008,068
11	埼玉県	さいたま市	7,204,168	3,798.08	1,896.79	7,194,556
12	千葉県	千葉市	6,211,820	5,156.61	1,204.63	6,216,289
13	東京都	新宿区	13,186,562	2,188.67	6,024.92	13,159,388
14	神奈川県	横浜市	9,059,616	2,415.86	3,750.06	9,048,331

データベース内容のXML API化

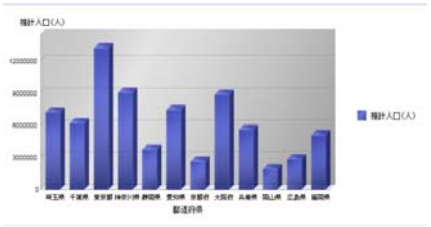
●各SQLテーブルの列名をxmlのタグにし，行を繰り返し要素にした形で，XML API形式で情報提供する．

・SQL DBをXML表示
Fuken_dbのxml
表示 xmlファイル生成？
Fukendataのxml
表示 xmlファイル生成？
Fukendata_nのxml
表示 xmlファイル生成？

```
<?xml version="1.0" encoding="UTF-8"?>
- <Fuken_db time="2012/11/07(水) 19:48:40">
  - <row num="1">
      <number>1</number>
      <tag1>北海道</tag1>
      <tag2>札幌市</tag2>
      <tag3>5,485,916</tag3>
      <tag4>78,420.86</tag4>
      <tag5>69.95</tag5>
      <tag6>5,506,419</tag6>
    </row>
  - <row num="2">
      <number>2</number>
      <tag1>青森県</tag1>
      <tag2>青森市</tag2>
      <tag3>1,363,006</tag3>
      <tag4>9,644.55</tag4>
      <tag5>141.32</tag5>
      <tag6>1,373,339</tag6>
    </row>
```

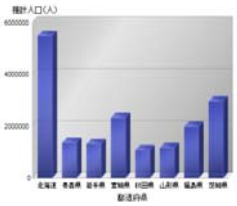

HTML5によるグラフ描画

●SQLテーブル表示で指定した列のデータを縦軸に，指定した行数分を横軸に表す．



HTML5によるグラフ描画

●描画用JavaScriptのflotr2ライブラリの指定形式に合わせ，SQLテーブルからデータを取り出し，JSを構成する．



```
<script type="text/javascript" src="excanvas.js"></script>
<script type="text/javascript" src="vbar.js"></script>
<script type="text/javascript">
window.onload = function() {
var g = new html5jp.graph.vbar("graph");
if( ! g ) { return; }
var items = [
["推計人口(人)", 5485916, 1363006, 1312756, 2323224, 1075055, 1161294, 1988995, 2956854, 2000021, 2000876]
 ];
 var params = {
  x: ["都道府県", "北海道", "青森県", "岩手県", "宮城県", "秋田県", "山形県", "福島県", "茨城県", "栃木県", "群馬県"],
  y: ["推計人口(人)"]
 };
 g.draw(items, params);
 }
</script>
```


適用性評価

(1)単純なテーブル構造，即ち<table>タグ配下に同じ数の<td>を持つ<tr>のみが並んでいる場合．列名とその配下のデータが一致する．
(2)(1)の構造で一部の行(<tr>)に列の不足がある．あるいは<td>の数は揃っているが，<td>の要素が「空」である．
(3)<tr>や<td>内で種々の属性が入っている．
(4)<tr><td>の間にスタイル指定などのタグが入っている．

※これらについては概ねOK．

適用性評価

(5)<table>の入れ子．即ち，テーブル内に何重かにテーブルが存在しているケース．
(6)各行で列要素が一定でない．(2)のケースで単純に補完が出来ないようなケース．特に，タイトル行が複数行あり小計的に区分され，データの要素との一致が取りにくい様な構造．
(7)HTML5のグラフ化で，系列や具体的数値の設定においてエラーを生ずることがある．

※完全に解決してはいない．

適用性評価

(5)ではテーブル構造検索の関数(メソッド)を再帰呼び出しにて取り出していくが，SQLテーブル構造に適切に変換できないことがあるなど，一部のみ適応可能である．

(6)は種々のケースがあり個々に対応しつつあるが，汎用的な解には至っていない．JavaのString型のreplaceFirst()メソッドなどを多用した実装で，このメソッドで対応できない文字，例えば「[,], *」などを見つけ次第エスケープしているが，今後も対応しきれていない箇所が出現する可能性は高い．

適用性評価

(7)についても，単純なケースは検証できているが，Flotr2ライブラリの限界を超える場合や，面積や人口というように単位の違うものの扱い（表示）に関して十分な対応ができていない．

●なお本機能の有用性の観点の主観評価では，研究室関係の被験者（20名）から全て良好との結果を得ている．

結論

●API化と情報提供がなされれば成長が期待できる分野は色々とある．

●HTMLの解析は困難であり，通常は二次利用が認められていないので，WebサービスAPIの提供が望まれる．

●サービスコンピューティングに向けたHTMLサイト利用基盤を構築した．