1ZF-07

オープンデータを授業利用するための スクレイピングWebAPIの開発

中智宏† 漆原宏丞† 本多佑希† 兼宗進†
†大阪電気通信大学

1 はじめに

現在インターネット上に公開されているデータはExcel、PDF、HTML、CSVなど様々な形式であり、それらの表のレイアウトも統一されていない。特に、行政が公開しているデータは、人間が画面や印刷物で見ることを目的に作られているため、縦や横のセルが結合されることが多く、機械的に表を切り出したときに、どこからどこまでが一つの表なのかを判別することが難しいことが多い。

このようなデータの例として、複数の表が一つに連結されている例を図1に、一つの表が無駄な空行で分割されてしまっている例を図2に示す。

そこで、PDFなどから切り出した表データを解析することで、どこからどこまでが一つの表であるのかを判別するシステムを開発した。システムの評価として、各都道府県が公開している新型コロナ陽性者のデータを解析する。

また、プログラミング授業での活用も期待することができる。その仕組みと授業での利用例も合わせて報告する。

36	15歳未満	Under	1503	770	733	15,105 7
37	15～64	years old	7466	3783	3683	74,710 3
38	65歳以上	and over	3619	1574	2045	36,079 1
39	うち75歳以	and over	1872	739	1133	18,657 7
40	うち85歳以	and over	620	196	423	6,112 1
41						
42					割 合	（単位 %）
43	15歳未満	Under	11.9	12.6	11.4	12.0
44	15～64	years old	59.3	61.7	57.0	59.3
45	65歳以上	and over	28.7	25.7	31.6	28.7
46	うち75歳以	and over	14.9	12.1	17.5	14.8
47	うち85歳以	and over	4.9	3.2	6.6	4.9

図 1: オープンデータの例1

35	（再 掲）	Regroupe				
36	15歳未満	Under	1503	770	733	15,105 7
37	15～64	years old	7466	3783	3683	74,710 3
38	65歳以上	and over	3619	1574	2045	36,079 1
39	うち75歳以	and over	1872	739	1133	18,657 7
40	うち85歳以	and over	620	196	423	6,112 1
41						
42					割 合	（単位 %）
43	15歳未満	Under	11.9	12.6	11.4	12.0
44	15～64	years old	59.3	61.7	57.0	59.3
45	65歳以上	and over	28.7	25.7	31.6	28.7
46	うち75歳以	and over	14.9	12.1	17.5	14.8
47	うち85歳以	and over	4.9	3.2	6.6	4.9

図 2: オープンデータの例2

2 スクレイピング

Web上のデータを抜き出する手法としてスクレイピングが存在する。このスクレイピングプログラムは、一般にはあるサイトのデータを抜き出す際には専用のプログラムを作る必要がある。サイトによって、HTML形式やJSON形式、CSV形式など提供されているデータの形式が異なるためである。そのため、汎用的に多くのデータを抜き出し、蓄積することは難しい。こういった取り組みは、これまでもいくつか行われている [1] [2]。

今回は、汎用的に多くのサイト等からデータをスクレピングする手法を検討した。対象としては、HTML(table)、PDF、エクセルを対象に表形式のデータを抜き出すことにした。これにより、学校での授業など学習の場において、Web上で提供されている多くのデータを用いた学習が可能になることを期待している。

3 本システムについて

3.1 概要

本スクレイピングツールは、WebAPIとして動作する。これにより、プログラミング言語によらない動作が実現できる。パラメータとしてスクレイピングを行うURLやその他の情報を渡すと、プログラムから扱いやすい、二次元配列の形式を始めとしたデータ形式で表データを取得することができる。図3に本スクレ

A data scraping Web API system for programming education

Yuji Tanaka†, Kousuke Urushihara†, Yuki HONDA†, Susumu KANEMUNE†
†Osaka Electro-Communication University
432-8011, Neyagawa, Japan

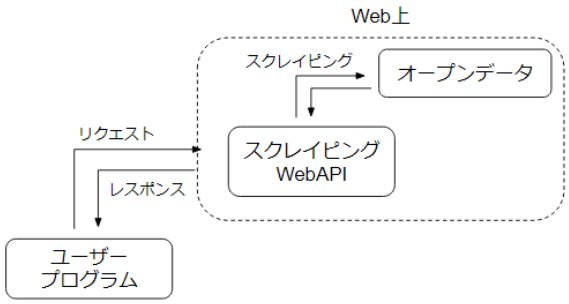

図 3: 本スクレイピングツールのシステムモデル

イピングツールのシステムモデルを示す。WebAPI として提供された本スクレイピングツールがユーザプログラムからリクエストを受けたら、パラメータとして渡された URL をもとにオープンデータにアクセスしてスクレイピングを行い、プロキシのような役割を果たす。

3.2 表範囲の自動検出

表データを自動で検出する手法を検討した。提供されているオープンデータは人が目で見るための形式であることが多く、プログラムから検出しやすい形にはなっていない場合が多い。そのため、プログラムから表データに書かれている各データが文字列なのか、数値なのか、または空白なのかを認識することにした。同じ型の行が複数行続いたら、その部分を一つの表データをみなす。この手法により、ある程度はデータから表を抜き出すことができると考えた。

オープンデータの一部を図 1、および図 2 に示す。これらのデータであれば、まず最初の行の場合、１列目が文字列、２列目が文字列、３列目以降は空白となる。２行目以降、数行に渡って１列目が文字列、２列目が文字列、３列目以降が数値という形式になっている。そのため、２行目以降の複数行を１つの表データとみなして抽出する。

3.3 各種形式への対応

入力データの形式によって、各データ形式に対応するデータ抽出モジュールを用意した。これらのモジュールにより、データから表を抜き出してから、再加工という形で表範囲を抜き出している。図 4 に、各モジュールの関係を示す。

4 評価

システムの評価として、各都道府県が公開している新型コロナウイルス陽性者のデータを解析した。４７都道府県が公開しているデータのうち、４２都道府県

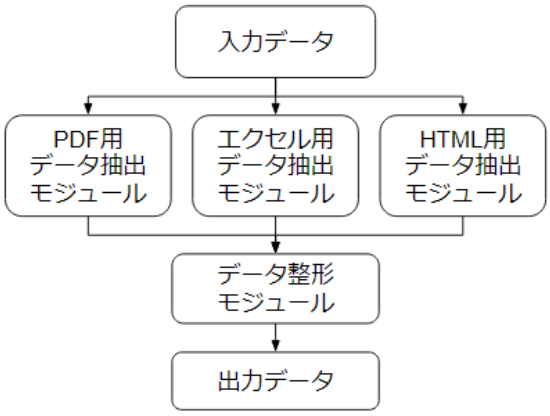

図 4: 各モジュールの関係

は自動でデータを判別することに成功した。自動判別が難しいケースへの対応は、今後も検討を進めていきたい。

5 授業利用の想定

今回のシステムは、例えばデータ分析の授業に活用することができると考えている。データ分析は重要な要素であるが、分析とする対象のデータの用意は教員にとって負担となる。そこで、本システムを利用することで、Web 上で公開されている多くのデータをそのまま授業で利用することができる。これにより、教員の負担軽減につながると考えている。

6 おわりに

Web 上のデータから自動で表データを抽出するスクレイピング WebAPI を開発した。今後は、授業で実際に利用することや、自動で定期的にデータの更新を確認しながら、データをサーバ上に蓄積するように拡張を行いたい。また、実際に学校の授業で活用するなど、今回開発したシステムの有効性についても確認したいと考えている。

参考文献

[1] 吉本龍司. 次世代ライブラリ：1. カーリル -図書館のオープンデータ化を促す仕組み. 情報処理, Vol.55, No.5, pp.446-451, 2014.

[2] 諏訪勇貴, 和田知華, 宇田隆哉. 形態素解析と機械学習を用いたオープンデータカタログサイトの集約手法. 情報処理学会研究報告, Vol2017-CSEC-76, No.1, pp.1-6, 2017.