ISS-P-214

Web スクレイピングを用いて既存システムを組み込んだプライベートクラウドの構築

田中　太樹†　　山内　雅弘††　　渡邊　敏正†††
† 近畿大学大学院 システム工学研究科 システム工学専攻　　†† 近畿大学 工学部 電子情報工学科
††† 広島大学大学院 工学研究科 情報工学専攻

1. はじめに

近年，様々な場面で現行システムからクラウドサービスへの移行が見受けられる．教育機関においても，乱立する数多くの現行教育系システムを廃止して，教育クラウドの新規導入が図られているが，クラウドを自前で一から構築するとなると負荷が大きい．既存システムをクラウドに組み込み継続利用できれば移行時の負担を軽減できる．API が公開されているシステムであれば API に沿ってプログラムを作成すればよいが，API が用意されていない場合には何らかの方法でシステムへアクセスする必要がある．そこで本研究では Web スクレイピング([1,2,3]等参照)に着目し，この技法を用いて既存システムを組み込み，継続利用を可能とするクラウドを試作した．

2. Web スクレイピング

Web スクレイピングとは，ユーザが欲する情報を(他サイトから)選択的に取得する方法のことで，目的の情報を持ったサイトから HTML を“取得”・“抽出”・“整形”することで実現される．Web スクレイピングを実現するための方法は多数存在するが，本研究では“取得”は PHP の cURL 関数(libcurl ライブラリ) [4]を利用し，“抽出”・“整形”はそれぞれプログラムを作成することで実現した．libcurl ライブラリは，多くの異なったプロコトルで様々なサーバと通信を行え，認証も多くの形式をサポートしている．

3. スクレイピング用スクリプトの実装と動作検証

クラウドの利用者はあくまでクラウドにアクセスするだけであり，既存システムのページへ直接アクセスさせない．既存システムのページ遷移等をクラウド上で忠実に再現するためには，取得した HTML のリンクなどを書き換える必要がある．そのために，(1)a タグ href 属性；(2)form タグ action 属性；(3)input タグ name 属性；(4)image タグなどの src 属性；の 4 種と，(5)それらがどこに書かれているか；の情報を検出する簡易 HTML パーサを作成した．検出した内容を基に，HTML に加工を施す．これと同時に，デザイン統一化などの整形やシングルサインオン実現のための処理を行うことで，既存システムをクラウドが提供する機能の一部のように見せることが可能となる．

なお，Web スクレイピングには[2]で述べられている技術面での問題だけでなく，悪用するとフィッシング詐欺になることや，元サイトの利用規約に抵触しかねないといった問題が存在するため，開発初期である現段階においては，学内で稼働している[6,7,8]のシステムを既存システムとして利用し，教育用途のプライベートクラウドを構築して動作検証を行った．[6,7]に関しては，所属研究室で過去に作成されたシステムで，API が用意されていない．また，[8]は moodle[5]ベースの LMS だが，あえて API を利用せずに Web スクレイピングを行い，本手法の検証を行った．

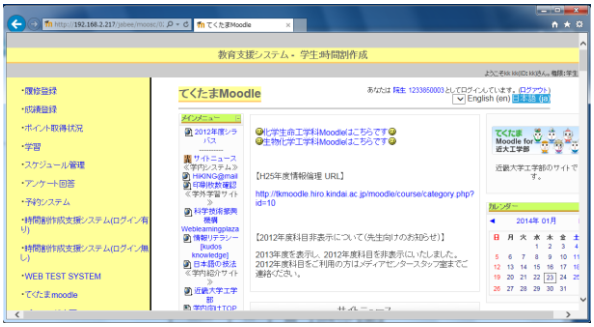

図 1　[8]へ Web スクレイピングを行った際の動作画面

紙面の都合上詳細は省略するが，図 1 は[8]へ Web スクレイピングを行い，ページを取り込んだクラウドの画面で，ページ遷移も問題なく行える．同様に，[6,7]のシステムでの動作も確認できた．

4. まとめと今後の課題

本研究では，(1)ベースとなるクラウドの作成；(2)Web スクレイピング用スクリプトの作成と実装；(3)既存システムを用いた動作検証；を行い，既存システムをクラウドで利用することができた．今後の課題として(1)クラウドの機能強化；(2)他の既存システムへの対応；などが挙げられる．

参考文献

[1] "WebScraping", http://www.sophia-it.com/content/Webscraping
[2] 西村紅美,塚本亨治,"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", IPSJ SIG Tech. Rep. Vol. 2009-SE-31, pp.73-80 (2009).
[3] 渡邊拓磨, "Web スクレイピングを用いた RFID システムによる企業間連携に関する研究", 法政大学大学院デザイン工学研究科紀要 Vol.2 (2013).
[4] "PHP:cURL", http://jp.php.net/manual/ja/book.curl.php
[5] "moodle", http://docs.moodle.org/2x/ja/
[6] 尺長義憲,東馬尚哉,"カリキュラム系統図と履修履歴を考慮した科目履修の対話的ナビゲーションシステムの試作", 近畿大学工学部卒業論文 (2009).
[7] 枝廣梢,小谷和広,百々勇輔,"ドラッグアンドドロップ形式の解答方法と問題難易度自動選択に対応したWeb試験システム", 近畿大学工学部卒業論文 (2009).
[8] "てくたまmoodle", http://tkmoodle.hiro.kindai.ac.jp/