1L-08

Ducky：単一 Web ページまたは階層構造をもつ大型 Web サイトからの情報抽出・統合・管理システム

金岡　慧†　　遠山　元道‡
慶應義塾大学理工学研究科情報工学専修†‡

あらまし

Ducky とは，Web ページの URL や CSS セレクタから成る独自のデータ抽出ルールに基づいて Web からデータ抽出を行い，XML，JSON，CSV 形式のファイルとして出力するシステムである．その独自のデータ抽出ルールの文法により，抽出されるデータのノイズ除去や，複数ページにまたがる大型サイトからのデータ抽出・統合なども可能なことが特徴である．本論文では *Ducky* の全体像を示す．

1. はじめに

Web は巨大な情報資源であり，これらを 2 次利用のために得たいというニーズがある．その場合，手作業で書き出すかスクレイピングプログラムを記述する方法が考えられる．前者の場合，手間と時間のコストが大きくなり，誤りが生じる可能性もある．後者の場合，知識のないユーザにとっては敷居の高い方法となる．そのため，Web ページやサイトを機械的に処理し，有用な情報を取り出す情報抽出技術の研究が現在までに数多く行われてきた[1]．中でも著者らは HTML 文書からデータを抽出し，構造化する Web ラッパーに注目した．Web ラッパーとは，Web ページから特定の情報を抽出する抽出ルールおよび抽出プログラムである．Web ラッパーはその抽出ルールによって HTML 文書から特定の情報の位置を把握し，抽出を行うことで，構造を再構成する．

Web ラッパーは自動生成，半自動生成など，自動のレベルの異なるものがこれまでに提案されてきた．自動生成の手法では，Web ページを入力としてページレイアウトを解析し，ページ内における抽出すべき情報の位置の把握，すなわち抽出ルールの導出を自動で行う必要がある．その際，事前に教師データとしてユーザにトレーニングサンプルを入力させるものも提案されてきた．Web ラッパーの自動生成，半自動生成とでは，抽出されるデータの精度とユーザに対する負荷といった点で，トレードオフが生じる．

これらの問題に対し，著者らは半自動の Web ラッパーを含む，Web ページからの情報抽出・管理を可能とする Ducky システムの提案，開発を行っている[2][3]．Ducky は様々な構造の Web ページ，複数のページから成る大型 Web サイトを対象にデータを抽出・統合することができるシステムである．

2. システム概要

Ducky の全体像を図 1 に示す．ユーザはシステムのブラウザ GUI を通して，データ抽出を行う Web ページを表示し，目的の要素をマウスオーバーで選択する．Web ページの URL と，目的の要素の位置情報はデータ抽出ルールに格納される．目的の要素の位置情報は，CSS セレクタで記述される．

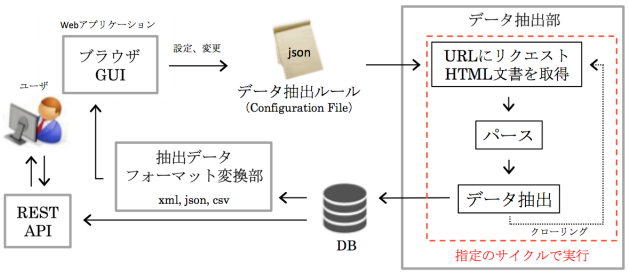

図 1　システム全体像

3. データ抽出

データ抽出に関する設定は，データ抽出ルールの *scraping* フィールドに定義される．そのフォーマットと定義されるフィールドを表 1 に示す．ブラウザ GUI を通してユーザによって選択された Web ページ上の目的の要素の位置情報は，CSS セレクタとして *scraping* フィールドに格納される．CSS セレクタは Xpath のように，マークアップ言語である HTML 文書の特定の部分を指定することができる言語構文である．図 2 の映画.com を例にとると，今回得たいデータが映画のタイトルとその公式 HP の URL の場合，それらは HTML の構造から全て同じパス，“div.unit li

Ducky : Data extraction, integration and management system for single Web page or hierarchically structured Web sites
† Kei Kanaoka Department of Information and Computer Science, Keio University
‡ Motomichi Toyama Department of Information and Computer Science, Keio University

表１　フィールド一覧

フィールド名	型	概要	
scraping	array	データを抽出するのに必要な以下のフィールドを定義	
url	string	起点となる URL を定義	
selector	string	現在のページから抽出したい要素の CSS セレクタを定義	
data	string	抽出されるデータをどのようなオプション (以下) で抽出するかを定義	
field	array	抽出データの項目名を定義	
attr	string	selector で指定したタグの属性を指定	
find	string	selector で抽出したタグの子要素を CSS セレクタで指定	
remove	array	“blank”	空白削除
		“parentheses”	括弧と括弧内の文字列の削除
		string	記述された正規表現，文字列を削除
replace	array	抽出されたデータの文字列置換	
next	object	起点のページからの遷移がある場合に使用	

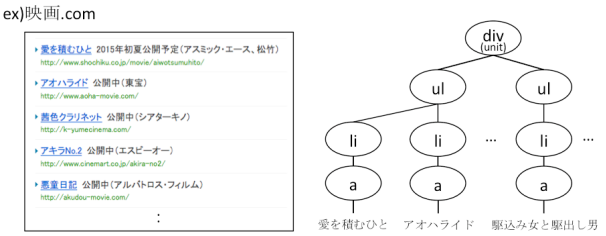

図２　HTML 構造

> a”で指定することが出来る．

設定ファイルで記述されたパラメータを元に，データ抽出部においてデータの抽出を行う．処理の流れは以下のようになる．

(1) “url”，“selector”の処理

指定された URL にリクエストを送り，HTML 文書を取得．指定された CSS セレクタを用いて取得した HTML 文書をパース．CSS セレクタで指定されたタグを要素として取得．

(2) “data”の処理

抽出したタグのテキストノードを抽　出　．“attr”や“find”が記述されている場合，抽出したタグの属性もしくは子要素のテキスト部分を要素として取得．“remove”や“replace”が記述されて いる場合，取得したデータに対して指定の処理を行う．

(3) next フィールドの有無

next フィールドが記述されている場合，取得した URL を “url”フィールドの値として処理(1)に戻る．記述されていない場合，取得したデータを DB へ渡す．“next”フィールドは同一テンプレートで生成されてい る遷移先 Web ページ群に対し，それぞれへリクエス トを送り，情報抽出・統合を行う．

4. アーキテクチャ

4.1 Web アプリケーション

ユーザはシステムのブラウザ GUI を通して，マウスオーバーやクリックの動作でデータ抽出を行うことができる．それらの動作はデータ抽出ルールとして登録される．抽出されたデータはブラウザ GUI または Web API を通して取得が可能となる．

4.2 Web API

Ducky では WEB API を通して抽出データを取得することができる．　実行したデータ抽出には，固有の file id がふりあてられる．また，ユーザは登録時にそれぞれ API key をもつ．この

```
/api/{FILE ID}?apiKey={USER API KEY}
```

file id と api key を使って，以下のように GET リクエストを送信すると，レスポンスとして最新の抽出結果が JSON 形式で返される．

5. 評価

提案システムの評価として，抽出されるデータの精度に関して既存の商用システムとの比較を行っている．

6. 結論

本論文では Ducky の全体像を示した．将来的には，ユーザによって抽出されたデータを，システムを利用するユーザ間で共有できるようにすることで，半自動ラッパーの Web ページごとにルール定義を行わなければならないという欠点を補っていきたいと考える．

文　献

[1] H.A. Sleiman and R. Corchuelo. A survey on region extractors from web documents. Knowledge and Data Engineering, IEEE Transactions on, 25(9):1960-1981, September 2013.

[2] Kei Kanaoka, Yotaro Fujii, and Motomichi Toyama. Ducky: A data extraction system for various structured web documents. In Proceedings of the 18th International Database Engineering & Applications Symposium, IDEAS’14, pages 342-347, ACM.

[3] Kei Kanaoka, Motomichi Toyama, Browser GUI for Generating Web Data Extraction Rules in Ducky, 17th International Conference on Information Integration and Web-based Applications & Services.