A performance of comparative study for semi-structured web data extraction model

Ily Amalina Ahmad Sabri, Mustafa Man
School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, Malaysia

ABSTRACT
The extraction of information from multi-sources of web is an essential yet complicated step for data analysis in multiple domains. In this paper, we present a data extraction model based on visual segmentation, DOM tree and JSON approach which is known as Wrapper Extraction of Image using DOM and JSON (WEIDJ) for extracting semi-structured data from biodiversity web. The large number of information from multiple sources of web which is image’s information will be extracted using three different approach; Document Object Model (DOM), Wrapper image using Hybrid DOM and JSON (WHDJ) and Wrapper Extraction of Image using DOM and JSON (WEIDJ). Experiments were conducted on several biodiversity website. The experiment results show that WEIDJ approach promising results with respect to time analysis values. WEIDJ wrapper has successfully extracted greater than 100 images of data from the multi-source web biodiversity of over 15 different websites.

1. INTRODUCTION
The extraction of the information from the large database is known as Knowledge Discovery (KD). Meanwhile data mining is the process of extracting the useful and relevant information from the database. It allows user to analyze data from different views and categorizing it, prior to concluding the relationships between data. The extraction and analysis of the web page is an interesting research area in the field of data mining and web mining. Internet has made the World Wide Web as the main pool for the collection and distribution of information to the users.

The reports from Internet World Stats states that there are now more than 4 billion of people around the world are using the internet (Stats, 2018). The latest data shows that Asia has become the biggest population of region contributing to more than 2 billion users. Figure 1 shows the world internet usage and population statistics. The number of internet users in earlier part of 2018 was 4,156,932,140 (Stats, 2018). This directly implies and contributes to tremendous growth of data in the World Wide Web.

Web data extraction is a technology developed over the past decade and encounters many new challenges. It has been discussed from different perspectives and views. It leverages on various scientific methods from various disciplines [1]. Laender, Ribeiro-Neto [2] Proposed taxonomy for data extraction approach to generate wrapper. Figure 2 shows the suggested classification proposed by the researchers. This classification is very important and helpful in order to understand the existing approaches for web data extraction.
Languages for wrapper development, TSIMMIS [3-5], WebOQL [6], Lorel [7] and Minerva [8] are some of the techniques that employ language for wrapper development. These approaches are used to address the problem of wrapper’s techniques. Various general languages were designed to construct wrapper. The example of languages that are used by programmers in developing wrappers are Perl and Phyton.

HTML aware, these approaches depend upon the structural features of web pages to perform both wrapper generation and data extraction. It is perform automatically without labour task. XWRAP, W4F and RoadRunner [9] are examples for this technique. Natural Language Processing (NLP) technique is used in order to mine facts from free text. Fact is indicated by entities and relationships between entities [10]. RAPIER [11], WHISK [12] and SRV [13] are some of the approaches of NLP. Wrapper induction, these approaches are based on certain features such as formatting. It can define the structure of data that found. The extraction rules can be generated based on training set. SoftMealy [14] and IEPAD [15].

Modelling based, these approaches try to find sections of the web pages that suit with pre-defined structures. NoDoSE [16] and DEByE [2] are the techniques for this category. Ontology based, this category is totally different with previous techniques. This is because it relying directly on the data. The obvious techniques for this approach are ODE [17] and DIADEM [18].

2. WEIDJ MODEL

In certain kind of web pages, it is common that data information of web pages are built dynamically according to specific template. Typical example is information of image, where image details has always the same structure, differing only in content loaded usually from a database. These data are rendered in the similar way. There have been discussed in former works [19-23] about how to extract data from web pages.

Figure 3 shows basic concepts of data extraction process. In preliminary step, user need to know what types of data that they want to extract either text, image, video or others. Then, they must decide which data need to be extracted. This selection must be done earlier because each data has their own source of type. Different data has their own source of data and relevant methods. After selection type of data done, next process could be proceed to abstract and transform selected data in tabular format using specific approach or methods. First, user need to understand the proof of concept for Web Data Extraction before develop a wrapper.
2.1. Problem formulation
As part of the input for the extraction, we suppose that the user has a number of structured web sources, denoted in the following, where each represents a set of web pages that describe images objects, which could be seen as relational tuples formed by three atomic type values; link, image, size. We assume a set of entity (atomic) types, where each such type represent an atomic piece of information. We continue by defining the typing formalism, by which any user can specify what data should be targeted and extracted from web page. We then describe the extraction problem.

2.1.1. Types and object description
In WEIDJ, it allows users to describe atomic types of objects. As building blocks for describing data, we assume a set of entity (atomic) types, where each such type \( t_i \) represents an atomic piece of information, expressed as images.

An instance of an entity type \( t_i \) is any images that exist. It is defined straightforward in a top-down approach and can be view as a tree structure whose internal nodes denotes the use of a complex type constructor. For example, image objects could be specified as a tuple type composed of three entity types: path of image, size of image, date of extraction and time of extraction. The first two entity types would be associated to predefined recognizers (for path of images and size of images), since this kind of information has easily recognizable representation patterns, while the last ones would have an instance of recognizers.

2.1.2. The extraction problem
For a given WEIDJ \( s \) and source \( S_I \), a template \( \tau \) with respect to \( s \) and \( S_I \) describes how instances of \( s \) can be extracted from \( S_I \) pages.

- For each set type \( t = [t_i, m] \) appearing in \( s \), \( \tau \) defines a separator string \( sep^t \), it denotes that consecutive instances of \( t_i \) will be separated by this string.
- For each tuple type \( t = \{t_1, ..., t_k\} \), \( \tau \) defines total images over the collection of types and a sequence of \( k + 1 \) separator strings \( sep_{t_1}, ..., sep_{t_k} \); this denotes that the instances of the \( k \) types forming \( t \), in the specified images, will be delimited by these separators.

The extraction problem can be described as follow. For a given input consisting of an WEIDJ \( s \) and a set of sources \( \{S_1, ..., S_n\} \),

1. set up type recognizers for all the entity types in \( s \),
2. for each source \( S_I \),
   a. find and annotate entity type instances or images in pages,
   b. infer a template \( \tau_I(s, S_I) \) based on the sample,
   c. use \( \tau_I \) to extract all the images of \( s \) from \( S_I \),
   d. select images that want to store in single multimedia database

A web page \( w \) is represented as a triple:

\[
 w = (b, S, R) \tag{1}
\]

Finite set of blocks is represented as:

\[
 b = \{w^1, w^2, ..., w^n\} \tag{2}
\]
All these blocks are not overlapped. Each block can be recursively viewed as a sub web page associated with sub structure inspired from the whole page structure. Finite set of separators such as horizontal and vertical are represented as:

$$S = \{s_1, s_2, ..., s_T\}$$  \hspace{1cm} (3)

Every separator in the same $S$ has same weight. Weight for each separator indicating its visibility. $R$ is the relationship of every two blocks in $b$. It can be expressed as

$$R = b \times b \rightarrow S \cup \{NULL\}$$  \hspace{1cm} (4)

For example, suppose $W_i$ and $W_j$ are two objects in $b$. $A = R(W_i, W_j) \neq NULL$ indicates that $W_i$ and $W_j$ are exactly separated by the separator $R(W_i, W_j)$. In other words, two objects are adjacent to each other, otherwise there are other objects between two blocks $W_i$ and $W_j$. Figure 4 shows layout structure visual segmentation of WWF web page while.

![Figure 4. Layout structure and visual segmentation of WWF web page](image)

3. RESULTS AND DISCUSSIONS

A key element, path of required HTML element is important which allows to locate and extract information. This section discuss the experimentation of data extraction for multi-uniform resource locator (URL). This experimental is different for data extraction from surface of web. This is because when working with multiple web of URL, user need to input several web URL that contain various information to be extracted from different structure of web pages. Figure 5 shows interface for multi-url.

The work described in this experimental work uses the same approach in the surface web. However, it involves more process and time consuming because the extraction process recursively traversing all element’s parents for each web page. Table 1 and Table 2 describes the benchmark of web address that are implemented on testing for image extraction of multiple web. Table 1 consists of three groups of URL such as group A, B and C. Each group has several web pages as shown in Table 2.
A performance of comparative study for semi-structured web data extraction

Figure 5. Interface for multi-sources of web

Table 1. Description of benchmark

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>General Biodiversity and Endangered Species Information</td>
</tr>
<tr>
<td>2</td>
<td>Databases and Datasets: Broad Scope</td>
</tr>
<tr>
<td>3</td>
<td>Databases and Datasets: Narrow Scope</td>
</tr>
</tbody>
</table>

Table 2. Details of benchmark

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Uniform Resource Locator (URL)</th>
</tr>
</thead>
</table>
| 1         | http://www.amnh.org  
http://www.ocean.si.edu  
http://www.iucn.org  
http://www.endangeredspeciesinternational.org  
http://www.wwf.org.my |
| 2         | http://www.gbf.org/  
http://www.unep-wcmc.org/  
http://www.naturereserve.org/  
http://www.organismnames.com/query.html  
http://www.catalogueoflife.org/col/search/all  
http://animaldiversity.ummz.umich.edu/site/index.html  
http://www.thplantlist.org/  
http://www.iucnredlist.org/  
http://www.itis.gov/  
http://www.consbio.org/  
http://bugguide.net/node/view/15740 |
| 3         | http://www.amphibiaweb.org/  
http://www.reefbase.org/  
http://primatelist.library.wisc.edu/  
http://bugguide.net/node/view/15740 |

Figure 6 shows the result for multi-sources of web pages for web data extraction using DOM, WHDJ and our proposed model, WEIDJ. This experiment has been tested on benchmark for three different group of URL that can be referred on Table 2. Time extraction is referred to the time loading for extraction process since the first image until the last image. The time extraction is measured in seconds. The time of WEIDJ retrieval on any data set is significantly lower than other methods and outperforms efficiently.
Table 3 (a, b) shows summarize results of extracting images on four different model of wrapper: DOM, WHDJ, WEID and WEIDJ no rules. As it can be seen from Table 3 (a, b), the result shows that the amount of image found are average similarly but the amount image that have been retrieved and filtered by each model are totally different. Images that retrieved using DOM approach are many compared to other models because the rules of filtering information are minimal. In addition, DOM approach does not consider the redundancy, It will retrieve the similar filename for each image if it exist in the multiple web address that acts as input. While page load or known as time extraction is used as primary benchmark to compare the performance for each model, the findings show that WHDJ is better that DOM because it uses JSON approach to transform data. JSON is approved that can decreased time better. While WHDJ is better than DOM, but we need to extract data in high speed. So we propose AJAX technology to make sure that the entire web page does not have to be reloaded each time the user requests a change during web data extraction.

Table 3(a). Results of images extraction for multiple web address (DOM, JSON)

<table>
<thead>
<tr>
<th>Group</th>
<th>Img found</th>
<th>Img retrieved</th>
<th>Img filtered</th>
<th>Time extraction</th>
<th>Img found</th>
<th>Img retrieved</th>
<th>Img filtered</th>
<th>Time extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>96</td>
<td>66</td>
<td>30</td>
<td>168.8032</td>
<td>86</td>
<td>54</td>
<td>32</td>
<td>133.2378</td>
</tr>
<tr>
<td>2</td>
<td>163</td>
<td>88</td>
<td>75</td>
<td>232.9429</td>
<td>143</td>
<td>78</td>
<td>65</td>
<td>184.783</td>
</tr>
<tr>
<td>3</td>
<td>94</td>
<td>27</td>
<td>67</td>
<td>121.7585</td>
<td>50</td>
<td>27</td>
<td>23</td>
<td>81.3768</td>
</tr>
</tbody>
</table>

Table 3(b). Results of images extraction for multiple web address (WEIDJ, WEIDJ (no-rules))

<table>
<thead>
<tr>
<th>Group</th>
<th>Img found</th>
<th>Img retrieved</th>
<th>Img filtered</th>
<th>Time extraction</th>
<th>Img found</th>
<th>Img retrieved</th>
<th>Img filtered</th>
<th>Time extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>86</td>
<td>43</td>
<td>43</td>
<td>18.1519</td>
<td>86</td>
<td>65</td>
<td>64</td>
<td>28.4313</td>
</tr>
<tr>
<td>2</td>
<td>129</td>
<td>65</td>
<td>64</td>
<td>28.4313</td>
<td>143</td>
<td>65</td>
<td>64</td>
<td>25.879</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>33</td>
<td>17</td>
<td>16.2757</td>
<td>50</td>
<td>33</td>
<td>17</td>
<td>15.6934</td>
</tr>
</tbody>
</table>

Figure 6. Performance for image extraction for multi-uniform resource locator (URL)

Table 4 shows the summary of time extraction in percentage. The percentage for each method shows the bigger comparison value especially for DOM approach. JSON is well known as fastest method [24]. That is the main reason why JSON can be fastest in extracting information compared to DOM method. Although JSON is good in performance of extracting data but the extracted information are not so efficient because this method will retrieve the same file image. So that, WEIDJ is proposed for this experimentation. WEIDJ is good in extracting beneficial information and the time performance is degrade. WEIDJ and WEIDJ no rules is the same method which apply combination of DOM and JSON but the difference between WEIDJ no-rules and WEIDJ method is WEIDJ no-rules will not filter noisy images. It will retrieve any images just like DOM and JSON.
4. CONCLUSION

We have proposed semi-structured web data extraction model, WEIDJ to extract data according to the predefined and simple data extraction rule. It consists of several parameters: link of the image which locates the data that has been retrieves from webpage, images, size of images and time extraction for each image. WEIDJ corresponds not only for single URL but also to multiple sources of web page. The current study limits its scope to extract images from surface of multiple web URL. Future research may consider image extraction from deep web of multiple URL.

ACKNOWLEDGEMENTS

I sincerely thank all those who helped me in completing this task especially Biasiswa Universiti Malaysia Terengganu (BUMT).

REFERENCES


Table 4. Results of images extraction in percentage for multiple web address (time extraction)

<table>
<thead>
<tr>
<th>Group</th>
<th>Time extraction</th>
<th>Percentage 100%</th>
<th>Time extraction</th>
<th>Percentage 100%</th>
<th>Time extraction</th>
<th>Percentage 100%</th>
<th>Time extraction</th>
<th>Percentage 100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>168.8032</td>
<td>50.41</td>
<td>133.2378</td>
<td>39.79</td>
<td>18.1519</td>
<td>5.42</td>
<td>14.6789</td>
<td>6.138</td>
</tr>
<tr>
<td>2</td>
<td>232.9429</td>
<td>49.35</td>
<td>184.783</td>
<td>39.15</td>
<td>28.4313</td>
<td>6.02</td>
<td>25.879</td>
<td>5.48</td>
</tr>
<tr>
<td>3</td>
<td>121.7585</td>
<td>51.79</td>
<td>81.3768</td>
<td>34.61</td>
<td>16.2757</td>
<td>6.92</td>
<td>15.6934</td>
<td>6.68</td>
</tr>
</tbody>
</table>

A performance of comparative study for semi-structured web data extraction ... (Iy Amalina Ahmad Sabri)
I. A. A. Sabri and M. Man, “Multiple types of semi-structured data extraction using WEID,” Regional Conference on Sciences, Technology and Social Sciences (RCSTSS), Universiti Teknologi Mara, Pahang: Copthorne Hotel Cameron Highlands, 2016.


BIOGRAPHIES OF AUTHORS

Ily Amalina Ahmad Sabri received her Diploma Information Technology from Polytechnic of Sultan Mizan Zainal Abidin (PSMZA). After that she enrolled to Universiti Malaysia Terengganu to further her degree studies in Software Engineering, which was obtained in 2009. She continued her master degree in Master of Science (Computer Science) in the same university and graduated in 2014. During masters’s degree, her research was in decision support system, focusing on FAHP in decision making for tourism destination. Now, she is a postgraduate student pursuing her Doctor of Philosophy (Computer Science), also in Universiti Malaysia Terengganu. Her current area of interest is web data extraction.

Mustafa Man is an Associate Professor in School of Informatics and Applied Mathematics and also as a Deputy Director at Research Management Innovation Centre (RMIC), UMT. He started his PhD studies in July 2009 and finished his studies in Computer Science from UTM in 2012. He has received Computer Science Diploma, Computer Science Degree, Masters Degree from UPM. In 2012, he has been awarded a “M Tec MOS Prestigious Awards” for his PhD by MIMOS Berhad. His research is focused on the development of multiple types of databases integration model and also in Augmented Reality (AR), android based, and IT related into across domain platform.