International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 6, December 2019, pp. 5463~5470
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i6.pp5463-5470

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

Article Info

Article history:

Received Jan 5, 2019
Revised Apr 18, 2019
Accepted Jun 12, 2019

Keywords:

Document object model (DOM)
Web data extraction
Wrapper extraction of image using DOM and JSON (WEIDJ)
Wrapper image using hybrid DOM and JSON (WHDJ)

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.

Corresponding Author:

Ily Amalina Ahmad Sabri,
School of Informatics and Applied Mathematics,
Universiti Malaysia Terengganu,
Kuala Terengganu, Terengganu Darul Iman, Malaysia.
Email: ilylina@yahoo.com

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 datebase. 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.

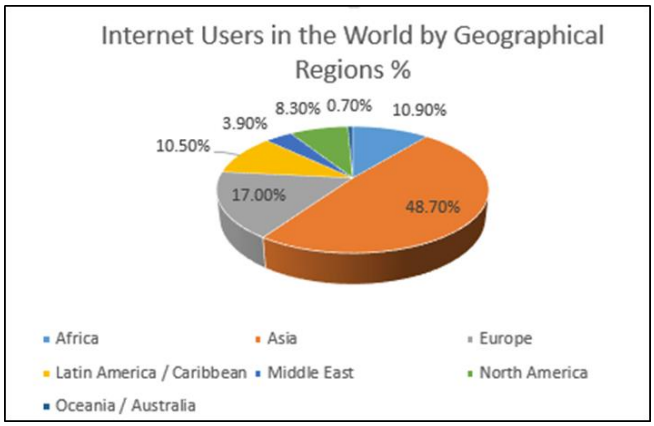

Figure 1. Number of internet users' for world wide web

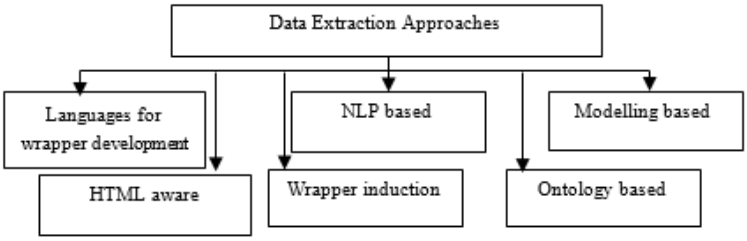

Figure 2. Data extraction approaches by Laender (2002)

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.

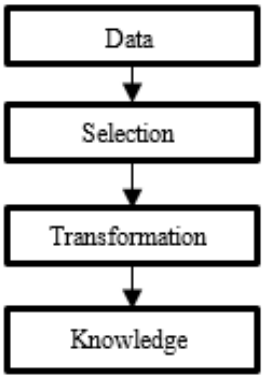

Figure 3. Basic data extraction process

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 τ with respect to s and S_i describes how instances of s can be extracted from S_i pages.

- For each set type $t = [\{t_i\}, m_i]$ appearing in s, τ defines a separator string sep^t, it denotes that consecutive instances of t_i will be separated by this string.
- For each tuple type $t = \{t_1, \dots, t_k\}, \tau$ defines total images over the collection of types and a sequence of $k+1$ separator strings $sep_1^t, \dots, sep_1^t$; this denotes that the k 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 $\{S_1, \dots, S_n\}$,

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 τ_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:

$$w = (b, S, R) \quad (1)$$

Finite set of blocks is represented as:

$$b = \{w^1, w^2, \dots, w^n\} \quad (2)$$

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:

$$S = \{s^1 s^2, \ldots, s^T\} \quad (3)$$

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

$$R = b \times b \rightarrow S \cup \{NULL\} \quad (4)$$

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.

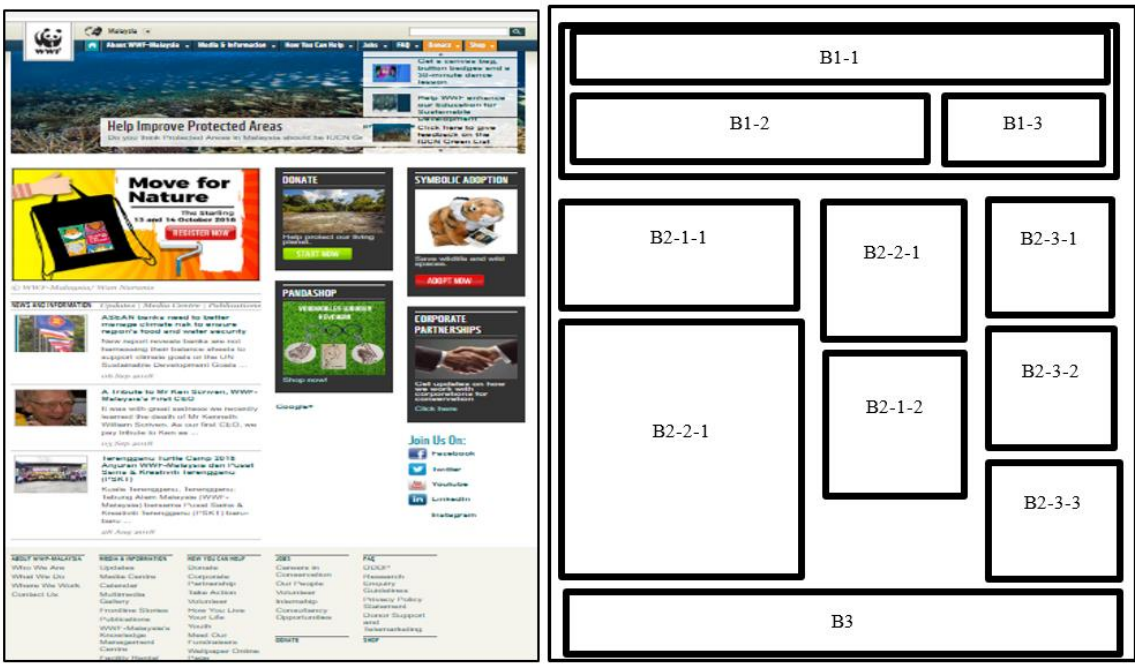

Figure 4. Layout structure and visual segmentation of WWF web page

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.

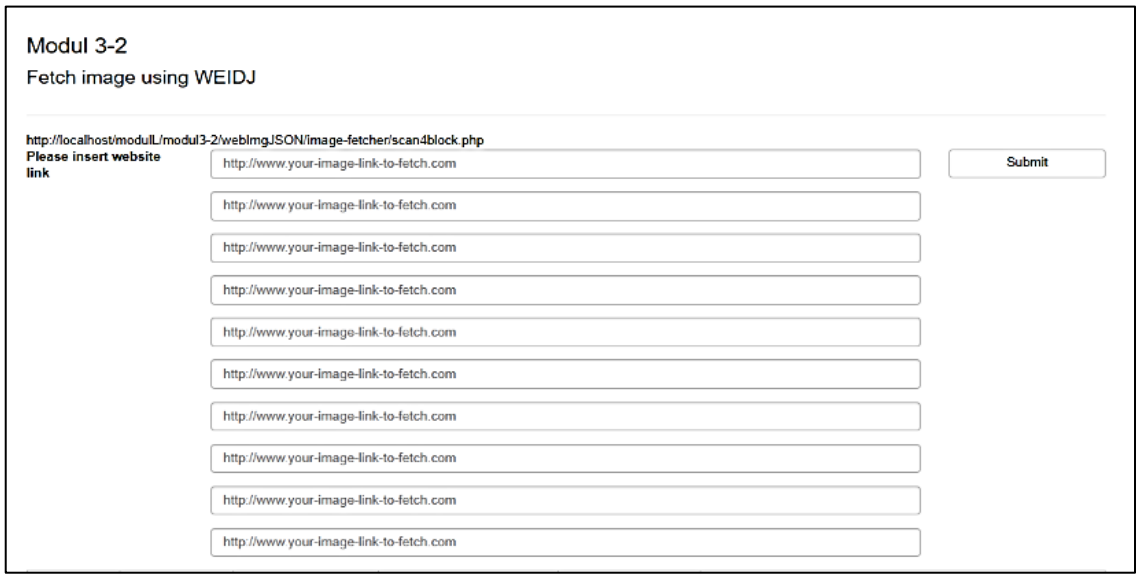

Figure 5. Interface for multi-sources of web

Table 1. Description of benchmark

Benchmark	Group of URL	Description
1	A	General Biodiversity and Endangered Species Information
2	B	Databases and Datasets: Broad Scope
3	C	Databases and Datasets: Narrow Scope

Table 2. Details of benchmark

Benchmark	Uniform Resource Locator (URL)
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.gbif.org/
	http://www.unep-wcmc.org/
	http://www.natureserve.org/
	http://www.organismnames.com/query.html
	http://www.catalogueoflife.org/col/search/all
	http://animaldiversity.ummz.umich.edu/site/index.html
	http://www.theplantlist.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://primatelit.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.

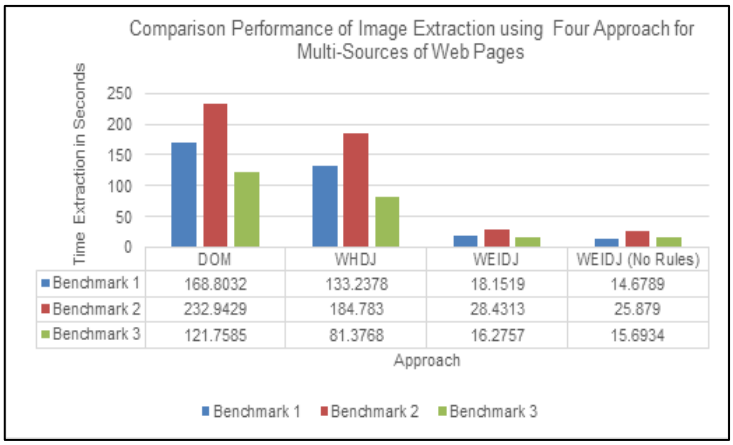

Figure 6. Performance for image extraction for multi-uniform resource locator (URL)

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)

Group	DOM				JSON			
	Img found	Img retrieved	Img filtered	Time extraction	Img found	Img retrieved	Img filtered	Time extraction
1	96	66	30	168.8032	86	54	32	133.2378
2	163	88	75	232.9429	143	78	65	184.783
3	94	27	67	121.7585	50	27	23	81.3768

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

Group	WEIDJ				WEIDJ (no-rules)	
	Img found	Img retrieved	Img filtered	Time extraction	Img retrieved	Time extraction
1	86	43	43	18.1519	86	14.6789
2	129	65	64	28.4313	143	25.879
3	50	33	17	16.2757	50	15.6934

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.

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

Group	DOM		JSON		WEIDJ		WEIDJ(no-rules)	
	Time extraction	Percentage 100%	Time extraction	Percentage 100%	Time extraction	Percentage 100%	Time extraction	Percentage 100%
1	168.8032	50.41	133.2378	39.79	18.1519	5.42	14.6789	4.38
2	232.9429	49.35	184.783	39.15	28.4313	6.02	25.879	5.48
3	121.7585	51.79	81.3768	34.61	16.2757	6.92	15.6934	6.68

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

[1] S. J. Nasti, *et al*., "A Comparative Study on Web Data Extraction Approaches," *International Journal of Engineering Science*, 2016.
[2] A. H. Laender, *et al*., "DEByE–data extraction by example," *Data & Knowledge Engineering*, vol. 40, pp. 121-154, 2002.
[3] J. Hammer, *et al*., "Extracting Semistructured Information from the Web," 1997.
[4] J. Hammer, *et al*., "Template-based wrappers in the TSIMMIS system," *ACM Sigmod Record*, 1997.
[5] J. Hammer, *et al*., "Semistructured Data: The TSIMMIS Experience," 1997.
[6] G. Arocena and A. Mendelzons, *"WebOQL: restructuring documents, databases and webs,"* International conference on data engineering, IEEE Computer Society, 1998.
[7] S. Abiteboul, *et al*., "The Lorel Query Language for Semi-Structured Data," *International Journal Digit Library*, vol. 1, pp. 68-88, 1997.
[8] R. Mooney, *"Relational learning of pattern-match rules for information extraction,"* Proceedings of the Sixteenth National Conference on Artificial Intelligence, 1999.
[9] V. Crescenzi, *et al*., "Roadrunner: Towards automatic data extraction from large web sites," *VLDB*, 2001.
[10] A. H. Laender, *et al*., "A brief survey of web data extraction tools," *ACM Sigmod Record*, vol. 31, pp. 84-93, 2002.
[11] M. E. Califf and R. J. Mooney, "Relational Learning of Pattern-Match Rules for Information Extraction," *AAAI/IAAI*, 1999.
[12] S. Soderland, "Learning information extraction rules for semi-structured and free text," *Machine learning*, vol. 34, pp. 233-272, 1999.
[13] D. Freitag, "Machine learning for information extraction in informal domains," *Machine learning*, vol. 39, pp. 169-202, 2000.
[14] C. N. Hsu and M. T. Dung, "Generating finite-state transducers for semi-structured data extraction from the Web," *Information Systems*, vol. 23, pp. 521-538, 1998.
[15] C. H. Chang and S. C. Lui, *"IEPAD: information extraction based on pattern discovery,"* Proceedings of the 10th international conference on World Wide Web, 2001.
[16] B. Adelberg, *"NoDoSE—a tool for semi-automatically extracting structured and semistructured data from text documents,"* ACM Sigmod Record, 1998.
[17] W. Su, *et al*., "ODE: Ontology-assisted data extraction," *ACM Transactions on Database Systems (TODS)*, vol. 34, pp. 12, 2009.
[18] T. Furche, *et al*., *"DIADEM: domain-centric, intelligent, automated data extraction methodology*," Proceedings of the 21st International Conference on World Wide Web, 2012.
[19] I. A. A. Sabri and M. Man, *"Multiple Types of Semi-structured Data Extraction Using Wrapper for Extraction of Image Using DOM (WEID),"* Regional Conference on Science, Technology and Social Sciences (RCSTSS 2016), N.A.Y.e. al., Editor, Springer Nature Singapore Pte Ltd., Singapore, pp. 67-76, 2018.
[20] I. A. A. Sabri and M. Man, "Improving performance of DOM in semi-structured data extraction using WEIDJ model," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 9, pp. 752-763, 2018.
[21] I. A. A. Sabri and M. Man, "The proposed algorithm for semi-structured data integration: Case study of Setiu wetland data set," *Journal of Telecommunication Electronic and Computer Engineering*, vol. 9, pp. 79-84, 2017.
[22] I. A. A. Sabri and M. Man, *"WEIDJ : An improvised algorithm for image extraction from web pages,"* The 8th international conference on information technology, Al-Zaytoonah University of Jordan (ZUJ), Amman, Jordan, IEEE Xplore, 2017.

[23] 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.
[24] Y. N. Li, *et al*., *"Mison: A Fast JSON Parser for Data Analytics,"* Proceedings of the Vldb Endowment, vol. 10, pp. 1118-1129, 2017.

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 "MIec 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.