Performance Analysis for Mining Images of Deep Web

Ily Amalina Ahmad Sabri[1], Mustafa Man[2]
Faculty of Ocean Engineering Technology and Informatics
Universiti Malaysia Terengganu
Terengganu, Malaysia

***Abstract*—In this paper, advancing web scale knowledge extraction and alignment by integrating few sources has been considered by exploring different methods of aggregation and attention in order to focus on image information. An improved model, namely, Wrapper Extraction of Image using DOM and JSON (WEIDJ) has been proposed to extract images and the related information in fastest way. Several models, such as Document Object Model (DOM), Wrapper using Hybrid DOM and JSON (WHDJ), WEIDJ and WEIDJ (no-rules) are been discussed. The experimental results on real world websites demonstrate that our models outperform others, such as Document Object Model (DOM), Wrapper using Hybrid DOM and JSON (WHDJ) in terms of mining in a higher volume of web data from a various types of image format and taking the consideration of web data extraction from deep web.**

Keywords*—Data extraction; Document Object Model; web data extraction; Wrapper using Hybrid DOM and JSON; Wrapper Extraction of Image using DOM and JSON

I. INTRODUCTION

As web sites are getting more complicated, the construction of efficient web information extraction systems becomes vital. A common arising issue in data extraction is the difficulties in locating the important segments, which contained the information being searched in a page. People always view a web page as different semantic objects rather than single objects. Some research efforts showed that the spatial and visual cues can help the users to unconsciously divide the web page into several semantic parts. Many recent works [1, 2] tried to extract the structured information from web pages using visual segmentations.

Deep web is known to contain more valuable informations compared to surface webs. Due to its complexities, deep web requires more efforts and the process may be time consuming. In addition, most of the web pages in deep web are generated only for visualizations, and the available data are not possible for exchange nor extraction. Thus, extraction process of deep web is crucial. Normally, each web page within a deep web, has similar characteristics; they share the same structures and templates. They are also encoded in a consistent way across all the pages. However, due to certain complexity issues, there are several deep web which are incompatible with the standards extraction process.

The primary drawback of the existing data extraction methods is the large volume of images to be extracted, making the process to be consuming. After the problem has been identified, the extraction process must be defined using the algorithms, either in pseudo code or flowchart to ensure the inputs, processes and outputs that will be involved in extraction process. The input is user's admittances such as web address. The process of image's extraction involves the construction of HTML documents into tree structures, finding the tags, checking of the images by segmentations and representation of the data in JSON format. The outputs are the extracted images and their related information, represented in tabular format.

There are some discussions of algorithms for data classifications, data extraction and data integration used in machine learning. They are also being used as the base algorithms in some data extractor systems. Researchers designed various models of algorithms to detect more extracted web contents and to increase the performance of extracting information. Besides that time processing for extracting data can be decreased.

Umamageswari, Kalpana [3] proposed Boyer Moore String Pattern Matching Algorithm. This algorithm was used to find patterns in the extracted data. The patterns were aligned with texts and then checked either it matched the opposing character of texts. The algorithm outperformed the efficiencies of other existing pattern matching methods.

Polynomial algorithm was proposed by Pouramini, Khaje Hassani [4] in their extraction algorithms. The page element was checked by traversing of the DOM tree in mixed approach; bottom up and top down. This polynomial algorithm nevertheless was only applicable for visible pages and cannot be applied for the whole web pages.

Heuristic algorithm was proposed for automatic navigation and information extraction from Scientific Publishers' Website [5]. It didn't require the processing of the entire DOM tree to identify the locations of attribute value pairs but it required linear time since it did not involve complex pattern matching or string alignment algorithms. Although the execution time was really reduced but it didn't work well in case of different domains such as structured journal pages linked to single publishers' websites.

Raza and Gulwani [6] discussed a predictive program synthesis algorithm for texts and web data extraction. The program synthesis technique for data extraction tasks included a general form of DSLs' extraction, a synthesis algorithm designed in a domain-parametric fashion, as well as concrete

DSLs and algorithm instantiations in the practical application domains of texts and webs extraction. Nonetheless, this predictive synthesis cannot be applied to other practical application domains such as richly formatted documents such as XML-based formats (DOCX, ODF, PPTX, etc.).

A method has been proposed to deal with deep web based on visual recognition combined with DOM analysis [7]. VRDE was proposed to solve the problems of DOM structures inefficiencies when dealing with large volumes of images and multi web pages [8]. The limitation of this research work was the presence of its noisy information which cannot be removed completely.

Manjaramkar and Lokhande [9] proposed DEPTA. This research extracted all the data based on the data mining regions, record extractions and record alignments. The proposed system used two major steps: the first step was the recognition of data regions inside the input web pages, following by its extraction. Second step was the arrangement of extracted data in structured formats. The tag substrings in record segmentation modules were examined and if there were any repetitions of tags, only one tag was selected. Partial tree alignment approach was based on tree matching, which means placing those records field in a couple of records that can be organized and liberated to other data field.

Kamanwar and Kale [10] agreed that Web Data Extraction is the process of extracting user's required information from web pages. Nowadays, the extractor is used to extract information because web page is an ocean of data which makes browsing information as a very complicated task. Normally the contents of web documents are unstructured. Web data extraction is defined as a process which use tool and wrappers as mediums to extract information from web documents in HTML format. The noisy information such as tags, advertisements, and banner will be removed by wrapper.

Fang [11] has proposed STEM to extract sequences of identifiers from the tag path of web pages. Then a suffix tree is built on top of these sequences and four refining filters are proposed to screen out the region which contains unnecessary information.

Pouramini, Khaje Hassani [4] proposed Handle-based Wrapper by using DOM tree approach. This research worked on text features as handles to extract data records from web pages such as textual delimiters, keywords, consents or text patterns. Polynomial algorithm has been proposed to check against the page elements in two situations; mixed bottom up and top-down traverse DOM-tree. Yet, the extraction process can only be performed on the visible parts, and not on the whole page, thus limits its further applications.

TANGO was presented by Jiménez and Corchuelo [12], aimed to learn rules for a precise and recallability extraction of information from semi-structured web documents. The high precision and recallability are pre-requisites in the context of Enterprise Systems Integration. It relies on an open catalogue of features that helps to map the input documents into a knowledge base in which every DOM node is represented by means of HTML, DOM, CSS, relational, and user-defined features.

II. Research Method

Wrappers are tools developed using specific techniques or models to be used for image's extraction from web. The wrapper is divided into two main parts. The first part involves the process of the insertion "URL" of web page. It involves the parsing of the HTML web page and storing them as Document Object Model (DOM) tree. The conversion from HTML web pages to DOM tree structure is important to understand the structure of HTML pages in tree environment. This method is useful in gauging the structure of data, whether it is structured, semi-structured or unstructured. The second part is related to the extraction techniques. The wrapper applies the extraction techniques; DOM, hybrid model of DOM and JSON (WHDJ) and hybrid model of DOM, JSON and visual segmentation (WEIDJ).

JSON is a syntax for storing and exchanging data. The advantage of JSON is an open-standard format that uses human readable text to transmit data objects [13]. The growing popularity of the JSON format has fueled increased interest in loading and processing JSON data within analytical data processing systems [14].

A JSON object is a key-value data format that is typically rendered in curly braces. When working with JSON, JSON objects can be represented in a *.json file*, but they can also exist as a JSON object or string within the context of a program. JSON may be coded with lots of lines, this shows that the format is generally set up with two curly braces (or curly brackets) that look like this { } on either end of it, and with key-value pairs populating the space between. Most data used in JSON ends up being encapsulated in a JSON object.

Key-value pairs have a colon between them as in "key": "value". Each key-value pair is separated by a comma, so the middle of a JSON looks like this: "key": "value", "key": "value", "key": "value". Fig. 1 shows the sample of output using JSON approach in key-value pairs.

A. DOM Tree

The Document Object Model (DOM) is a programming API for HTML and XML documents. People can create and build documents using DOM [15]. Besides that this model can be used to manipulate elements and contents of HTML and XML documents such as add, modify or delete [13, 16].

Page level data extraction system is developed using DOM Tree by Narawade, Prabhakar [17]. There are two types of technique for data extraction; online and offline. There are three stages involves; web page renderers, section selectors, and pattern generators. A web page renderer will accept a *url* as inputs from users. After that the web page will be displayed in web browser. Then, DOM tree structure will be applied for content extraction and its representation in structured formats. The section selector acts to divide web pages into different sections that enable user to select particular record or whole data section. The system will dynamically extract the contents from the different structured web pages such as blogs, forums, articles, etc. Pattern generator creates relatively absolute patterns based on the origin of selected data regions.

Fig. 1. Output Presentation of JSON.

B. Wrapper using Hybrid DOM and JSON (WHDJ)

This section describes the extraction techniques of web data using hybrid DOM and JSON approaches. Images are extracted by converting HTML documents into tree structure and the engine will try to find the location of tags for each image. Then, JSON approach will be applied in order to retrieve and transform image's information into an array.

C. Wrapper Extraction of Image using DOM and JSON (WEIDJ)

Finding and extracting the complete images from web pages without failure is complex since a special operation is needed to find tags and extract all images without missing any values. This study presents Wrapper Extraction of Image using DOM and JSON (WEIDJ) [13]. This method can solve this problem by converting the HTML documents to a tree structure and composing web pages into several sections based on visual segmentations as shown in Fig. 1. Then, the engine will try to find tags for each image. This approach also considers certain rules to be filtered before the image extraction process. The rules will encompass the filtering of noise information, repetitive image filenames and non-related images to the website or web page. After the process of filtering noise information, the extracted images will be transformed into structured format also known as tabular form. The image's filenames will be indexed before storing them into multimedia database.

D. WEIDJ No-Rules

This method has same flows to WEIDJ model but WEIDJ no-rules will retrieved images without filtering noise images nor considering the repetition of file images. However, the performance of extraction process in terms of extraction time is still the best compared to DOM and WHDJ.

Fig. 2 shows an example on visual segmentation of layout structure for www.wwf.org.my. A visual segmentation is developed using each leaf node as an object. Visual segmentation is proposed because it is easier to be understood as compared to texting in details. This segmentation is important to check the availability of required information in each block. During the experimental works, problems of data extraction have been detected. There are certain images that cannot be extracted. As the solution, the availability of images in each block should be checked. The advantage of structuring partition is it can provide a user friendly view for the web page [13].

At this level, all suitable visual blocks contained in the current web page will be recognized. Basically, every node in the DOM tree can be presented as a visual block but nodes such as <TABLE> and <P> are not suitable to be represented as a single visual block. Several rules in extracting the visual blocks are as below[13]:

- Tags cue such as <hr> usually displayed as a horizontal rule in visual browser. If each DOM node contains this tag, the section will be divided.
- If a DOM node has different background colour from one of its child node, it will not be divided into any segments.

When appropriate blocks are extracted, the rest of invalid nodes will be ignored. Separator can be used as indicator to divide different section within a page. These visual blocks segmentation is applied to check every single multimedia element so that all required information can be retrieved [13].

The advantages of this proposed wrapper is user can automatically select all images to store in single multimedia database or select by manually. Fig. 3 and Fig. 4 show the examples of selecting images either by automatically or manually.

Fig. 2. Layout Structure and Visual Segmentation of WWF Web Page.

Fig. 3. Select Images with Automatic Selection.

Biasiswa Universiti Malaysia Terengganu (BUMT)

Fig. 4. Select Images with Manual Selection.

Fig. 5. Biodiversity Dataset.

III. Results and Discussions

The rapid development of computer and technologies has increased the usage of web. Deep web is a part of webs. It contains various useful information, which are beneficials to public users. The contents is dynamically generated from various data sources. The numbers of web pages for deep web is extremely large. Thus extraction of images from deep webs is very complicated and time consuming.

The dataset called Science and Technology Resources on the Internet, "Biodiversity Web Resources" has been selected, which consists of 43 online databases [18] as shown in Fig. 4 and Setiu Wetlands web page, namely, WWF-Malaysia [13]. The benchmark of datasets that have been applied in this experimental work are from [18] [19] for instant dataset. A web page, WWF- Malaysia was used for the testing of real datasets as shown in Fig. 5. The two categories of these datasets are dense. The overall characteristics of these benchmark dataset are tabulated in Table I.

The overall characteristics of these benchmark dataset are tabulated in Table I. There are three domain that have been selected; General Biodiversity and Endangered Species Information, Databases and Datasets: Broad Scope and Databases and Datasets: Narrow Scope. There are seventeen uniform resource locators that have been used in mining images from deep web.

WWF stands for World Wide Fund for Nature. It was formerly known as the World Wildlife Fund but the name was later changed to indicate it also deals with environmental issues rather than wildlife alone [16]. Fig. 6 shows main page for WWF website that contains images, text, symbol logo and others.

Data storage and data extraction industries are the basis of large data analysis and mining. The storage and extraction big data is relatively simple but become complicated when dealing with large volumes of data. Web page contains many valuable information. The extraction of semi-structured data, especially those involving large images, is limited by the its long execution times. For this reason, crawlers play important role to solve this problem. In this research, experiments were performed based on extracting images from the whole website.

TABLE I. Characteristics of Instant Dataset

URL	**Uniform Resource Locator (URL)**	
General Biodiversity and Endangered Species Information		***Domain***
1	http://www.amnh.org/	American Museum of Natural History (AMNH) Hall of Biodiversity
2	http://ocean.si.edu/	Ocean Portal: Smithsonian Institution
3	http://www.iucn.org/	International Union for Conservation of Nature
4	http://www.endangeredspeciesinternational.org	Endangered Species International
5	http://www.wwf.my	World Wide Fund for Nature
Databases and Datasets: Broad Scope		***Domain***
6	http://www.gbif.org/	Global Biodiversity Information Facility (GBIF)
7	http://www.unep-wcmc.org/	UN Environment Programme: World Conservation Monitoring Centre (UNEP-WCMC)
8	http://www.natureserve.org/	NatureServ
9	http://www.organismnames.com/query.htm	Index of Organism Names: ION
10	http://www.catalogueoflife.org/col/search/all	Catalogue of Life
11	http://animaldiversity.ummz.umich.edu/site/index.HTML	Animal Diversity Web
Databases and Datasets: Narrow Scope		***Domain***
14	http://bugguide.net/node/view/15740	BugGuide.Net
15	http://www.amphibiaweb.org/	AmphibiaWeb
16	http://www.reefbase.org/	ReefBase
17	http://primatelit.library.wisc.edu/	PrimateLit

Fig. 6. World Wide Fund for Nature (WWF) Web Page.

Normally, information in deep web is presented in "divs" sections. In the experiment, 30 pages or URLs have been selected randomly from the same website. The time taken for extraction of one page to another page was calculated. From the deep web result page, this research shows that WEIDJ is efficient because it can retrieve and extract semi-structured data from level of pages. The result shows that this model is very effective and can be used to extract the data quickly and accurately. The problems related to extraction's efficiencies and accuracies of different deep web page heterogeneity have been solved. The interference of web page noise [7] to data extraction also can be removed completely.

Fig. 7 shows the results of the crawl, the abscissa represents the number of pages extracted and the vertical axis represents the total time taken to extract the corresponding pages of WWF.

Table II shows the result of experimental of image extraction for web crawler. There are seven benchmarks that have been considered in this experimentation such as; 'Link Found', 'Img found', 'Img retrieved', 'Img filtered', and 'Time'. The analysis shows that the total of link found is in within range between three models but the findings of image retrieved is different between these models. Due to their genericities, the existing unsupervised approaches have significant drawbacks. The proposed alternative technique is more suitable especially to be applied in real life scenarios and offers several advantages WEIDJ are enabling to mix different types of images. As we know, there are various type of images that can be found in web pages such as TIFF (also known as TIF, file ending with .tif), JPEG (also known as JPG, file ending with .jpg), GIF (file ending with .gif), PNG (file ending with .png) and raw image files. Types of recognizers for each image is very important as the extraction process can be improved using the structural features of data. Fig. 8 shows types of image that can be retrieved by WEIDJ model.

Fig. 9 shows the output for data extraction by corresponding page. The extraction time will be calculated from the beginning of the extracted page to the next page [16].

The description of targeted images allows user to avoid the extraction of unnecessary images such as logos and buttons. Fig. 10 shows the example of images that have been extracted by WHDJ approach. This approach will ensure the extraction of useful information and all the noisy information and images will be neglected.

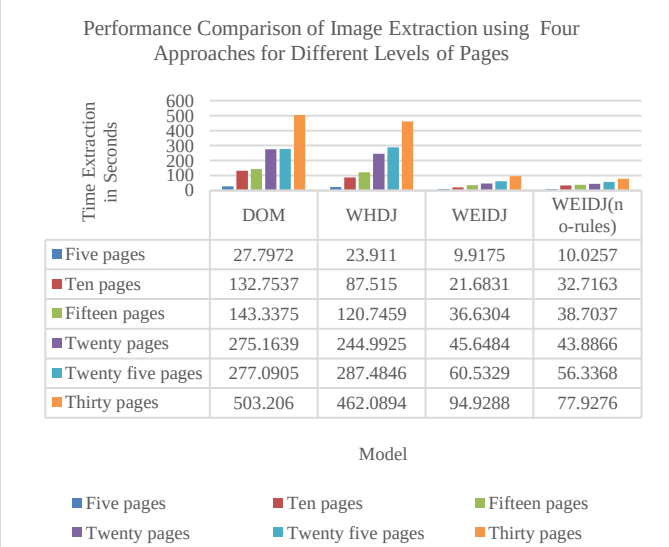

Fig. 7. The Extracted Pages of WWF Website.

```
if(empty($imagetype)) return false;
switch($imagetype)
{
    case 'image/bmp': return '.bmp';
    case 'image/cis-cod': return '.cod';
    case 'image/gif': return '.gif';
    case 'image/ief': return '.ief';
    case 'image/jpeg': return '.jpg';
    case 'image/jpg': return '.jpg';
     case 'image/pipeg': return '.jfif';
    case 'image/tiff': return '.tif';
    case 'image/x-cmu-raster': return '.ras';
    case 'image/x-cmx': return '.cmx';
    case 'image/x-icon': return '.ico';
    case 'image/x-portable-anymap': return '.pnm';
    case 'image/x-portable-bitmap': return '.pbm';
    case 'image/x-portable-graymap': return '.pgm';
    case 'image/x-portable-pixmap': return '.ppm';
    case 'image/x-rgb': return '.rgb';
    case 'image/x-xbitmap': return '.xbm';
    case 'image/x-xpixmap': return '.xpm';
    case 'image/x-xwindowdump': return '.xwd';
    case 'image/png': return '.png';
    case 'image/x-jps': return '.jps';
    case 'image/x-freehand': return '.fh';
    default: return false;
}
```

Fig. 8. Types of Image.

TABLE II. IMAGE EXTRACTION BY WEB PAGES FOR BENCHMARK

General Biodiversity and Endangered Species Information																	
Model	DOM					WHDJ					WEIDJ					WEIDJ(no-rules)	
Benchmark	Link Found	Img found	Img retrieved	Img filtered	Time	Link Found	Img found	Img retrieved	Img filtered	Time	Link Found	Img found	Img retrieved	Img filtered	Time	Img retrieved	Time
amnh.org	132	4881	2125	2756	10556.2238	132	4013	2028	1985	8778.3747	132	1521	1385	136	457.7495	4934	475.1597
ocean.si.edu	97	1966	1610	356	2319.4244	97	1705	1505	200	2076.7548	96	836	803	33	312.985	2011	329.4163
iucn.org	96	999	811	188	1979.6851	96	707	596	111	1625.4596	96	340	310	30	308.6347	2008	278.9528
endangeredspeciesinternational.org	6	77	44	33	150.9678	6	59	44	15	91.347	30	262	102	160	26.4048	411	25.9593
wwf.org.my	**142**	**1803**	**1374**	**429**	**1900.9394**	**142**	**1626**	**1370**	**256**	**1385.7157**	**143**	**1311**	**1059**	**251**	**318.2913**	**1899**	**285.2288**

DATABASES AND DATASETS: BROAD SCOPE

Benchmark	Link Found	Img found	Img retrieved	Img filtered	Time	Link Found	Img found	Img retrieved	Img filtered	Time	Link Found	Img found	Img retrieved	Img filtered	Time	Img retrieved	Time
http://www.gbif.org/	62	342	106	236	545.4321	62	205	101	104	361.3751	62	143	107	36	147.4583	603/357	142.1595
http://www.unep-wcmc.org/	30	310	227	83	695.9005	30	247	216	31	573.0375	30	195	189	6	129.1575	312/304	114.6983
http://www.natureserve.org/	63	767	557	210	929.8449	63	626	498	128	791.9172	64	347	274	73	77.2083	822/802	97.894
http://www.organismnames.com/query.htm	21	270	113	157	377.3488	21	223	111	112	304.9616	34	282	280	2	35.9359	2145/720	50.1937
http://www.catalogueoflife.org/col/search/all	33	314	8	306	363.7471	33	316	80	236	371.8078	43	130	228	229	65.0781	1896/509	62.099
http://animaldiversity.ummz.umich.edu/site/index.html	79	631	589	42	1020.3173	79	510	474	36	757.0506	79	383	257	126	163.663	125	108.732
http://www.theplantlist.org/	32	606	266	340	700.2953	32	457	219	238	528.9477	32	220	115	105	42.9643	942/598	53.3552
http://www.iucnredlist.org/	92	4536	3290	1246	8610.6563	92	4081	3280	801	6222.5125	95	2760	2551	209	432.0595	4614/4591	440.4091
http://www.itis.gov/	39	465	108	357	1253.9414	39	167	107	60	559.0691	40	70	32	38	78.4108	479	70.2567
http://www.consbio.org/	39	374	231	143	204.4485	39	342	215	127	290.8351	40	141	131	13	96.2886	383	107.2494

DATABASES AND DATASETS: NARROW SCOPE

Databases and Datasets: Narrow Scope																	
Model	DOM					WHDJ					WEIDJ					WEIDJ(no-rules)	
Benchmark	Link Found	Img found	Img retrieved	Img filtered	Time	Link Found	Img found	Img retrieved	Img filtered	Time	Link Found	Img found	Img retrieved	Img filtered	Time	Img retrieved	Time
http://bugguide.net/node/view/15740	83	5100	4695	338	9843.5567	58	1218	1018	200	2561.8086	88	5292	3914	1317	1040.6848	8632/5100	1396.5794
http://www.amphibiaweb.org/	9	127	50	77	167.6131	9	27	21	6	55.1403	3	77	69	8	4.6227	158/79	6.6687
http://www.reefbase.org/	95	5276	119	5157	4071.2581	95	1196	119	1077	757.0451	99	206	133	73	178.8349	5478/5532	173.0226
http://primatelit.library.wisc.edu/	7	15	5	10	45.5	7	15	5	10	44.6401	12	29	14	15	34.7044	29/14	39.8865

Fig. 9. Extracting 5 Pages.

Fig. 10. Example of Noisy Images.

IV. Conclusion

The World Wide Web contains large unstructured and semi-structured data. Researchers are welcomed to develop and implement various techniques to extract data from web sources due to the need for structured information. A wide range of web data extraction in several fields has been developed and continues to be proliferated. In this paper, WEIDJ as the best approaches among several models; DOM and WHDJ have been discussed. This wrapper provides three different level of extraction for DOM, WHDJ and WEIDJ. This paper discusses about extracting images for deep web using DOM, WHDJ and WEIDJ models. It is easy for human to extract images by just entering the data (web URL for extraction) without any needs to be informed of the instructions involved during the extraction process. The proposed method has numbers of advantages over previous extraction based techniques. In the first part of this paper, the applications of web data extraction systems to real world scenario were reviewed. The focus is how the application can work in practices and classify two models: DOM and JSON that have been applied by discussing several models. A simple implementation was provided in extracting multimedia data focusing on image using several websites. In future work, we plan to extend our approach to extract data from multi-web page. The performance of image extraction will influence the time for extraction process. Proposed technique gives good result in time processing in extracting data. The impact of the study for the nation building is the extraction of image that can be used for other purposes.

Acknowledgment

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

References

[1] Malhotra, P. and S.K. Malik. Web page segmentation towards information extraction for web semantics. in International Conference on Innovative Computing and Communications. 2019. Springer.

[2] Gulati, P. and M. Yadav, A novel approach for extracting pertinent keywords for web image annotation using semantic distance and euclidean distance, in Software Engineering. 2019, Springer. pp. 173-183.

[3] Umamageswari, B., R. Kalpana, and V. Archana, Web Data Extraction System using Boyer Moore String Pattern Matching Algorithm. 2018.

[4] Pouramini, A., S. Khaje Hassani, and S. Nasiri, Data extraction using content based handles. Journal of AI and Data Mining, 2017.

[5] Kumaresan, U. and K. Ramanujam, Web Data Extraction from Scientific Publishers' Website Using Heuristic Algorithm, 2017.

[6] Raza, M. and S. Gulwani, Automated Data Extraction using Predictive Program Synthesis. 2017.

[7] Cai, Z., et al., A Vision Recognition Based Method for Web Data Extraction, 2017.

[8] Cai, D., et al., VIPS: a vision-based page segmentation algorithm. 2003, Microsoft Technical Report, MSR-TR-2003-79.

[9] Manjaramkar, A. and R.L. Lokhande. DEPTA: An efficient technique for web data extraction and alignment. in Advances in Computing, Communications and Informatics (ICACCI), International Conference on. 2016. IEEE.

[10] Kamanwar, N. and S. Kale. Web data extraction techniques: A review. in Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), World Conference, 2016. IEEE.

[11] Fang, Y.X., et al., STEM: A suffix tree-based method for web data records extraction. Knowledge and Information Systems, 2018. vol. 55(2), pp. 305-331.

[12] Jiménez, P. and R. Corchuelo, On learning web information extraction rules with TANGO. Information Systems, 2016. vol. 62, pp. 74-103.

[13] Sabri, I.A.A. and M. Man, Improving performance of DOM in semi-structured data extraction using WEIDJ model. Indonesian Journal of Electrical Engineering and Computer Science, 2018. vol. 9(3), pp. 752-763.

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

[15] Sabri, I.A.A. and m. Man, A performance of comparative study for semi-structured web data extraction model. International Journal of Electrical and Computer Engineering (IJECE), 2019. vol. 9(6), pp. 5463-5470.

[16] Sabri, I.A.A., et al. Web data extraction approach for deep web using WEIDJ. in 16th International Learning & Technology Conference 2019. Riyadh, Saudi Arabia: Elsevier.

[17] Narawade, S.M., et al., A web based data extraction using hierarchical (DOM) tree approach. International Journal for Innovative Research in Science and Technology, 2016. vol. 2(11), pp. 255-257.

[18] Sabri, I.A.A. and M. Man, Improving performance of DOM in semi-structured data extraction using WEIDJ model. Indonesian Journal of Electrical Engineering and Computer Science, 2018. vol. 9(3), pp. 752-763.

[19] Creech, J. Biodiversity web resources. Science and Technology Resources on the Internet 2012 [cited 2017 31 May 2017]; Available from:http://www.istl.org/12-fall/internet.html.