A deep web data extraction model for web mining: a review

Ily Amalina Ahmad Sabri, Mustafa Man
Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, Kuala Nerus, Terengganu, Malaysia

Article Info

Article history:

Received Sep 11, 2020
Revised Apr 22, 2021
Accepted May 1, 2021

Keywords:

Data extraction techniques
Document object model
Noisy information
Web data extraction
Wrapper extraction of image using DOM and JSON
Wrapper using hybrid DOM and JSON

ABSTRACT

The world wide web has become a large pool of information. Extracting structured data from a published webpages has drawn attention in the last decade. The process of web data extraction (WDE) has many challenges, due to variety of web data and the unstructured data from hypertext markup language (HTML) files. The aim of this paper is to provide a comprehensive overview of current web data extraction techniques, in terms of extracted quality data. This paper focuses on study for data extraction using wrapper approaches and compares each other to identify the best approach to extract data from online sites. To observe the efficiency of the proposed model, we compare the performance of data extraction by single web page extraction with different models such as document object model (DOM), wrapper using hybrid dom and json (WHDJ), wrapper extraction of image using DOM and JSON (WEIDJ) and WEIDJ (no-rules). Finally, the experimentations proved that WEIDJ can extract data fastest and low time consuming compared to other proposed method.

Corresponding Author:

Ily Amalina Ahmad Sabri
Faculty of Ocean Engineering Technology and Informatics
Universiti Malaysia Terengganu
Kuala Nerus, Terengganu, Malaysia
Email: ilylina@umt.edu.my

1. INTRODUCTION

The World Wide Web has become a large pool of information which contains web pages, including images, audio, video clips, product information. Web traffic is among the important issue due to the extraction process [1]. The process of extracting data from web pages is a concern for people that lead to other purpose and give huge benefit. Commonly, websites are mainly design for human to glance certain information. The structure of websites are different each other and they are semi-structured. People need to select certain images mmanually that they are interested to save. It is time consuming. One of the technologies that can be applied for web data extraction (WDE) is called as a wrapper. The main goal of this wrapper or tool is to transform the semi-structured data into structured data. There are a lot of researches that discuss about wrappers. Most researches discuss about automatic data extraction includes noise information. A post-processing may be required in web data extraction to deal with beneficial extraction. It is important to extract the data with high precision and recall and also in fastest way for users. In this paper, a wrapper has been proposed to extract data based on different rules and models such as document object model (DOM), wrapper using hybrid DOM and JSON (WHDJ), wrapper extraction of image using DOM and JSON (WEIDJ) and WEIDJ (no-rules). This research works not only focus on how to extract data but also focus on providing user friendly platform for developers to treat the extracted data. This can be achieved completely through the user friendly browser for GUI.

Over the past new decades, numerous studies have been carried out on mining data from website or web pages and numerous techniques have been applied [2]. Many recent works tried to extract the structured information from web pages using variety of techniques such as DOM, visual segmentations or other techniques [3], [4]. Kamanwar *et al*., [5] agreed that WDE is a way of mining user's requisite figures 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 complex 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 hypertext markup language (HTML) format. The noisy information such as tags, advertisements, and bannerx will be removed by wrapper.

STEM has been proposed by Fang [6] to extract structures 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 view the sections which contain unnecessary information. Pouramini *et al*, [7] proposed handle-based wrapper by using DOM tree approach. This research worked on text features. It acts as handles to mine data records from web pages. The extraction consists of textual delimiters, keywords, constants or text patterns. Polynomial algorithm has been designed to form against the page elements in two situations; mixed bottom up and top-down traverse DOM-tree. The limitation of this application is the extraction process can only be performed on the visible parts. It can not extract from the whole web pages.

TANGO was proposed by Jiménez *et al*., [8], designed to learn rules for a detailed 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 depends on on an open catalogue of types that helps to map the contents of documents into a knowledge base. Each component of web documents in DOM node is denoted by HTML, DOM, CSS, relational, and user-defined features. Research done by *et al*., [9] has proposed the deep web data extraction (DWDE) framework to provide accurate results to users based on their URLs or domains searched.

Tripathy *et al*., [10] proposed VEDD wrapper to extract the relevant search results records (SRRs) from search engine by filtering out the noisy and redundant records. BFS was used in the beginning as it helped to re-structure the unstructured and semi-structured SSR pages which simplify the extraction process. SSR pages which in turn simplifies the extraction process. Derouiche *et al*, [11] proposed object runner technique called wrapper inference that processes the extraction and integration automatically of complex structured data. The extraction process was done in two stages; automatic annotations and extraction template constructions.

XWRAP, a wrapper based on DOM tree was developed by Liu *et al*., [12]. It consists of four components; syntactical structure normalization, information extraction was used for deriving rules, code generation was used for generating the wrappers programs, testing and packing used for validation. OLERA was developed by Chang *et al*., [13]. It produced extraction rules from semi-structured web pages without considering the training datas. It was designed with visualization supports. However, the technique was represented by its sensitivity to the ordering information. There were also probabilities in the failure of extraction process, if templates for each attribute were similar.

Liu *et al*., [14] proposed MDR. It was a fully automated system to identify data records in webpages. The application of this technique obliged all data to have same parents and multiple data records to have similar structures. The drawback of this approach was its disability to extract individual fields. VIPS was proposed by Cai Yu *et al*., [15] and Cai *et al*., [16]. It was a combination of two techniques; parsing of HTML in DOM tree and web page layout analysis using visual cues. The experiments clearly showed that vision-based web page content structure was very helpful in detecting and filtering out noisy and irrelevant information. Although this research proved good compliances to the multiple data regions of deep webs for data extraction, it still restricted by its incapability it completely removing noise.

Crescenzi *et al*., [17] developed RoadRunner. This tool enabled data extraction through the use of automatically generated wrappers. It was based on the similarities and differences between the webpages. The advantage of Road Runner is that it had no prior knowledge about the schema of the webpages and its ability in handling nested structures of contents. The limitations were its disability in managing disjunction cases and errors in the input documents, thus affecting it's effectiveness. IEPAD, a system that automatically discovered extraction rules from web pages [18]. This system can identify record boundaries from repeated pattern mining and multiple sequence alignments. The advantage of this technique is the extraction of information involves no human efforts and content dependent heuristics. The limitation of this tool was its poor ability in dealing with complex and nested structured data.

Hsu *et al*., [19] developed SoftMealy as web data extraction tool. This tool applied contextual rules and finite state tranducers (FST) technique which comprised body tranducers and tuple transducer. The body tranducers extracted the parts of the web contents that contain tuple. Then, tuple tranducers iteratively extracted the tuples. This technique however was not able to generalize overseen separators. TSIMMIS was an extractor that extracts data using extractor from WWW contents then converted the extracted information into a structured format before storing it into database [20]. The relevant data is retrieved in object exchange model (OEM) format.

Web data extraction system is a software application that can retrieve relevant information such as text, images, audio and many others from web sources [21]. This application usually cooperates with web sources and mining the relevant information to be stored. The mining contents consist of origins in the HTML web pages and can be post-processed, transformed to the most suitable structured format and stored for advance purpose. DOM can be applied directly to discover the required information from HTML documents. Abidin *et al.,* [22] constructed DOM tree structure on the preliminary step. Then, unnecessary nodes such as script, style need to be filtered. Classification process is vital to the search classes of multimedia data. Data for media will be recognized when the parser found word “src=” in the data structure. Finally multimedia data can be extracted. However, it has been found that large amount of processing times are required for the extraction of web pages which consists large size of HTML structures. Besides that, all images will be extracted without considering repetitive files. Thus WEIDJ model is proposed to overcome the limitations of DOM model in extracting images. Table 1 summarizes web data extraction tools.

The motivation for this research originates from previous works on techniques and methodologies of locating and extracting data from various web pages of different sites. These data can be very beneficials and useful for managerial information. The extracted information is merged into the multimedia database and can be used to fulfill new queries in the next stage of data mining. The main contribution of this research work is the development of the web data extraction model using hybrid approaches for images extraction and details revealation of its information. This model is expected to enables an effective image’s extraction by specifically disclose only related parts, simultaneously results in a reduced extraction’s times. This paper is structured as follows; In the following Section 2, this paper presents the research method to address the extraction issues. Then, we will show the performance of proposed tool in Section 3 which presents result and analysis and finally in Section 4, the conclusion is discussed.

Table 1. Web data extraction tools

(Author,year)	Tools	Model
Fang, Xie, Zhang, Cheng and Zhang [6]	STEM	Suffix Tree Based Method
Pouramini, Khaje Hassani and Nasiri [7]	Handle-based Wrapper	DOM Tree
Jiménez and Corchuelo [8]	TANGO	DOM
Chitra and Aysha Banu [9]	DWDE	Tag based Feature
Tripathy, Joshi, Thomas, Shetty and Thomas [10]	VEDD	- DOM Tree - Breadth First Search (BFS)
Derouiche, Cautis and Abdessalem [11]	ObjectRunner	
Liu, Pu and Han [12]	XWRAP	DOM Tree
Chang and Kuo [13]	OLERA	
Liu, Grossman and Zhai [14]	MDR	
Cai, Yu, Wen and Ma [15]	VIPS	- DOM Tree - Visual Cues
Crescenzi, Mecca and Merialdo [17]	Road Runner	
Chang and Lui [18]	IEPAD	Pattern Discovery
Hsu and Dung [19]	SoftMealy	
Hammer, Garcia-Molina, Cho, Aranha and Crespo [20]	TSIMMIS	Object Exchange Model (OEM)

2. RESEARCH METHOD

The basic concepts of data extraction process must consist of data, selection, transformation and knowledge. In the preliminary step, users need to know the types of data that they are extracting either texts, images, videos or others. This selection of data must be done earlier because each data has their own sources and extracting models. After the selection of the type of data has been done, the following process are abstracting and transforming the selected data into tabular format using specific approaches which need to be fully understood prior to develop a wrapper.

Wrappers are tools that have been developed using specific techniques or models. This tool can be used to extract images automatically. The wrapper can be categorized into two main components. The first component involves the insertion of web address,“URL” of web page. It comprises the parsing of the HTML web page and converting them to DOM tree structure. This conversion is significant to understand the structure of HTML pages in tree environment. This method is useful in handling the structure of data, whether it is structured, semi-structured or unstructured. The second part is related to the knowledge based construction. The extraction techniques that are been applied in this research work are DOM, hybrid model of DOM and JSON (WHDJ) and hybrid model of DOM, JSON and visual segmentation (WEIDJ). Figure 1 shows general models for three web data extraction models; DOM [23], WHDJ[24] and WEIDJ [25].

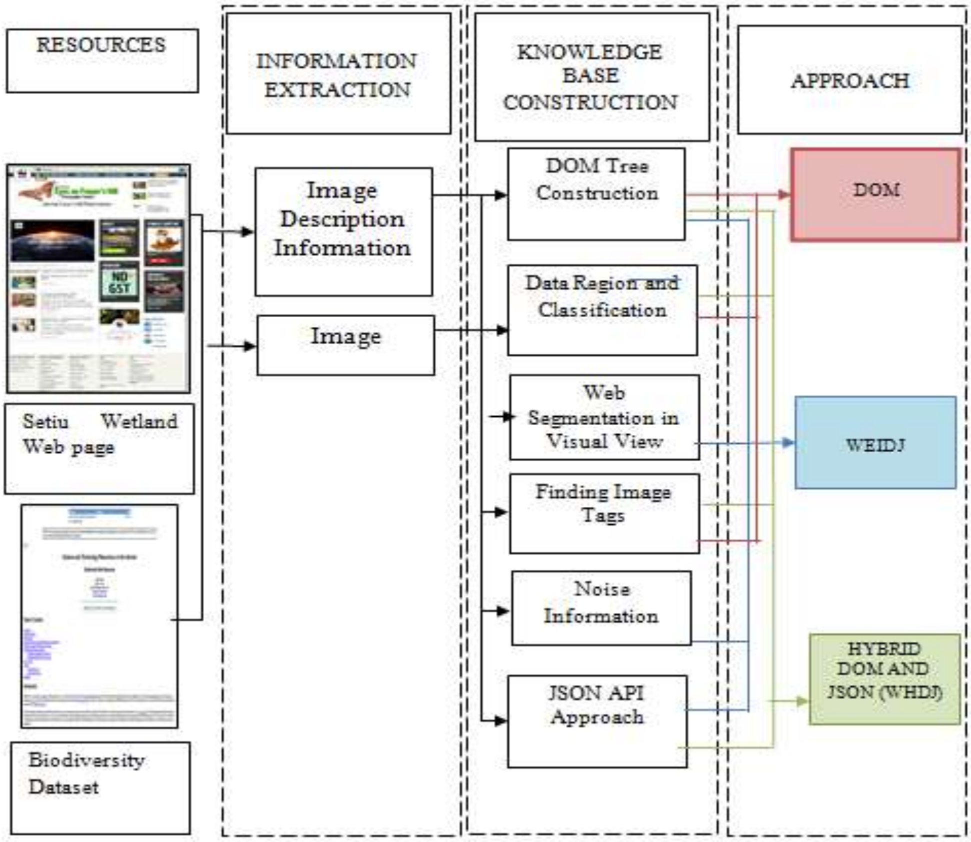
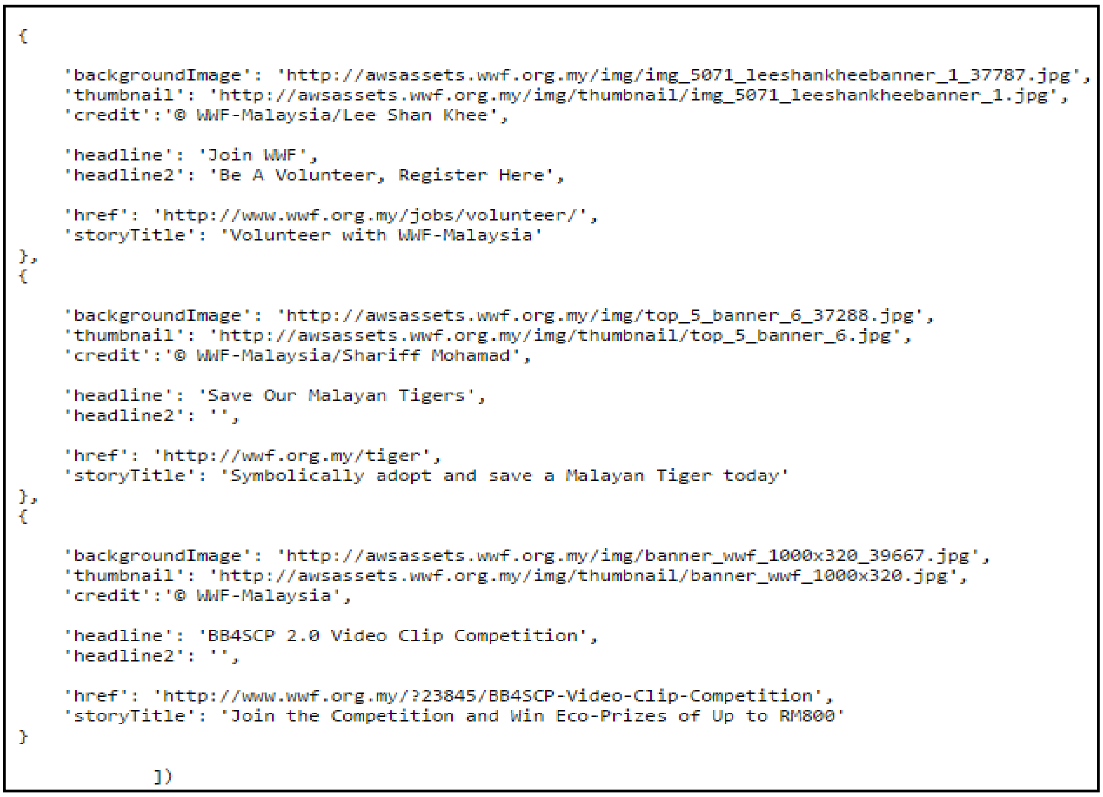


Figure 1. General model

3. RESULTS AND ANALYSIS

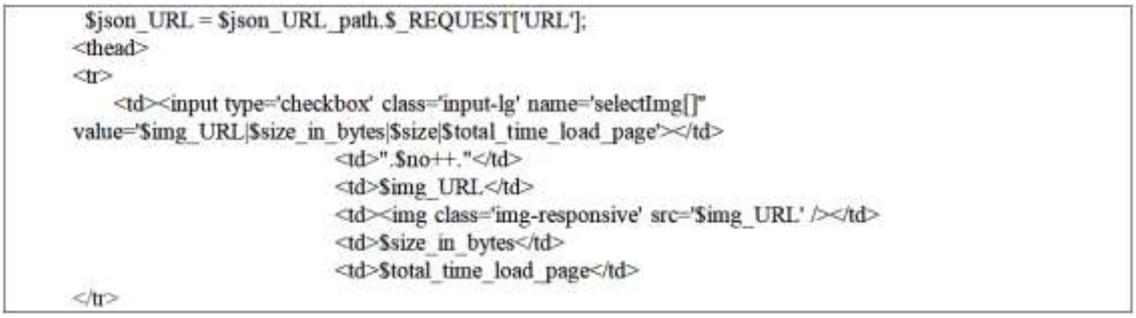
For experimental works, several samples of WWF web pages were taken and the extraction of contents was performed on the sampled data using HTML source files. This file contains the information of the images which are going to be extracted. Figure 2 shows sample extractor specification of file. List of commands consist of images and image's URL can be seen in the brackets '{' and '}'. Most of the images are in .jpg format file. WEIDJ is capable in extracting images of various formats such as .jpg, .gif, .bmp and others.

JavaScript Object Notation, also known as JSON is syntax for saving and swapping data. JSON has more benefit that can impress the user. This technology enables users easy to understand and get the the important text in order transmitting data objects. It is the best choice for storage and it also enabled a speedy response to information queries. The output can be ranged from simple to complex structure and highly nested. $json_url_path is used as constructor to inform the JSON data set to include the nested structures of JSON object. In first step, URL needs to be declared as json path. Then, 'src' value needs to be specified as path. This is very important in order to find the information of images from the image nested structure. Figure 3 shows the structure of extracted information, which has been organized in structured ways and displayed in table formats [26]. The extraction process in this example was performed by table definitions. The initial command $json_URL gets the contents of the source file or web address whose *URL* is given in ['*URL*']. After the file has been fetched, the contents will be detailed into specific criteria such as $no, $img_URL, image, $size_in_bytes and $total_time_load_page. The extraction information will be denoted in table representation.



```
{
    'backgroundImage': 'http://awsassets.wwf.org.my/img/img_5071_leeshankheebanner_1_37787.jpg',
    'thumbnail': 'http://awsassets.wwf.org.my/img/thumbnail/img_5071_leeshankheebanner_1.jpg',
    'credit':'© WWF-Malaysia/Lee Shan Khee',

    'headline': 'Join WWF',
    'headline2': 'Be A Volunteer, Register Here',

    'href': 'http://www.wwf.org.my/jobs/volunteer/',
    'storyTitle': 'Volunteer with WWF-Malaysia'
},
{
    'backgroundImage': 'http://awsassets.wwf.org.my/img/top_5_banner_6_37288.jpg',
    'thumbnail': 'http://awsassets.wwf.org.my/img/thumbnail/top_5_banner_6.jpg',
    'credit':'© WWF-Malaysia/Shariff Mohamad',

    'headline': 'Save Our Malayan Tigers',
    'headline2': '',

    'href': 'http://wwf.org.my/tiger',
    'storyTitle': 'Symbolically adopt and save a Malayan Tiger today'
},
{
    'backgroundImage': 'http://awsassets.wwf.org.my/img/banner_wwf_1000x320_39667.jpg',
    'thumbnail': 'http://awsassets.wwf.org.my/img/thumbnail/banner_wwf_1000x320.jpg',
    'credit':'© WWF-Malaysia',

    'headline': 'BB4SCP 2.0 Video Clip Competition',
    'headline2': '',

    'href': 'http://www.wwf.org.my/?23845/BB4SCP-Video-Clip-Competition',
    'storyTitle': 'Join the Competition and Win Eco-Prizes of Up to RM800'
}

            ])
```

Figure 2. A sample extractor specification file



```
      $json_URL = $json_URL_path.$_REQUEST['URL'];
    <thead>
    <tr>
        <td><input type='checkbox' class='input-lg' name='selectImg[]'
    value='$img_URL|$size_in_bytes|$size|$total_time_load_page'></td>
                              <td>".$no++."</td>
                              <td>$img_URL</td>
                              <td><img class='img-responsive' src='$img_URL' /></td>
                              <td>$size_in_bytes</td>
                              <td>$total_time_load_page</td>
    </tr>
```

Figure 3. The extracted information in JSON format

In addition to the basic capabilities of WEIDJ, our extractor also provides several other useful and user's friendly features. One of them is the queries to the saved images are provided. Figure 4 shows collection of images that have been saved in single multimedia database. These images can be queried from database for beneficial purpose. Thus, it can be used for further purpose such as generation of reports, analysis.

Images that have been selected will be stored in multimedia database. The images are succesfully saved in database. There are two options that can be selected by users for saving images into multimedia database either in automatic or manual. JSON as a standard module could accept any data structure and turn them into a representation of string. Figure 5 shows images that succesfully extracted and represent in JSON format. The advantages using JSON is faster and it is very easy to use.

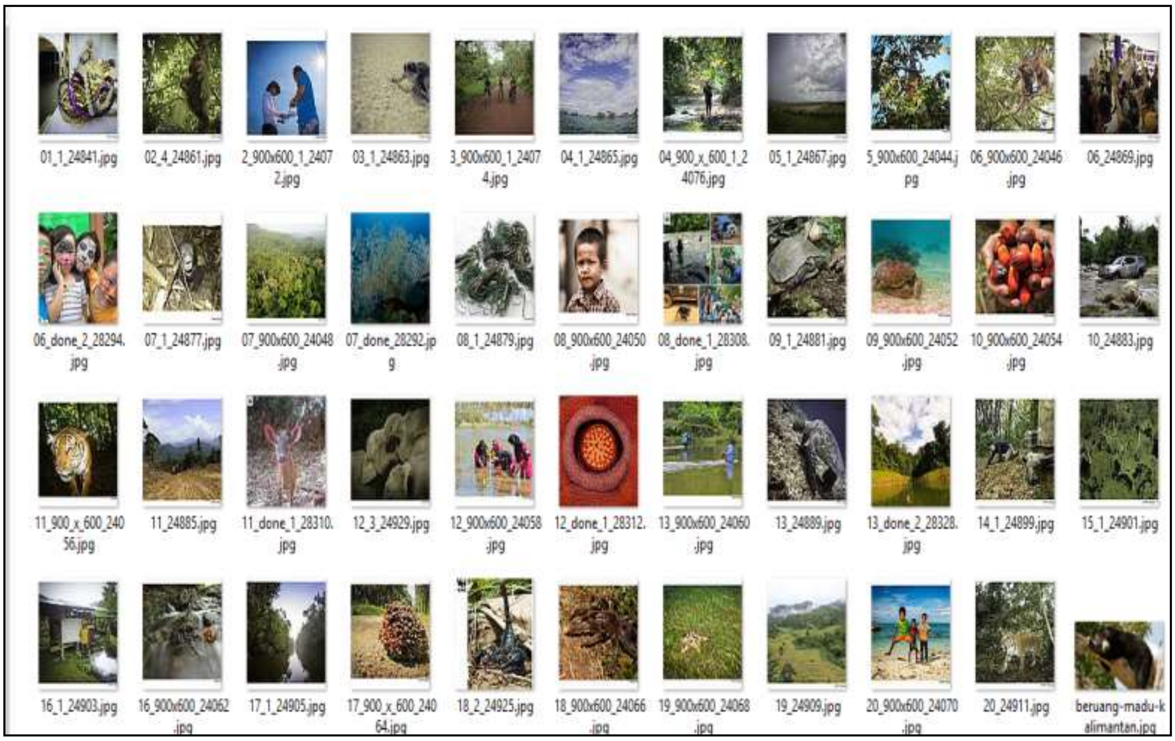
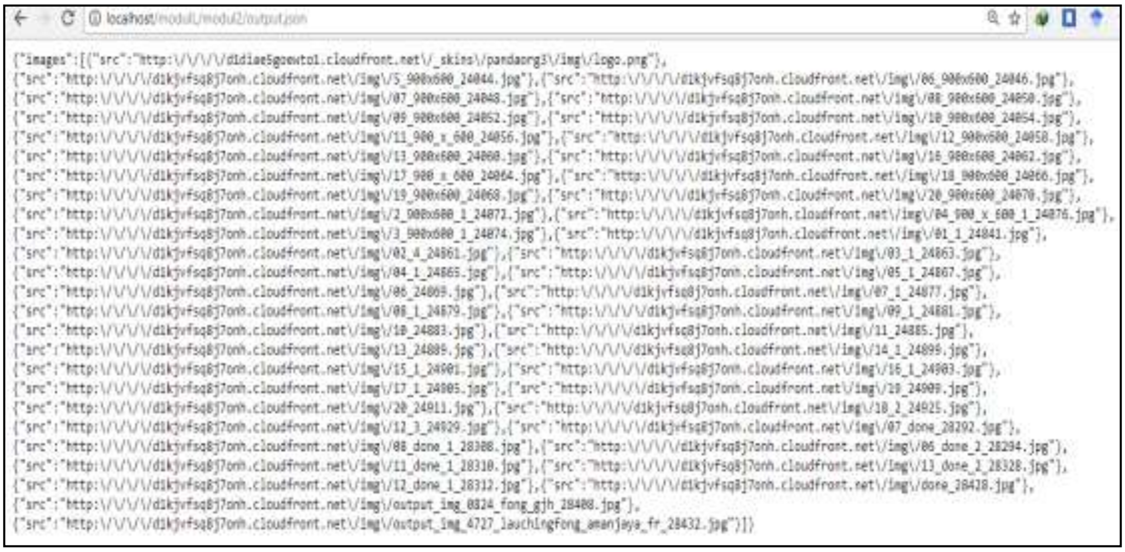


Figure 4. Images retrieved from database



```
{"images":[{"src":"http:\/\/\/\/d1diae5goewto1.cloudfront.net\/_skins\/pandaorg3\/img\/logo.png"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/5_900x600_24044.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/06_900x600_24046.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/07_900x600_24048.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/08_900x600_24050.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/09_900x600_24052.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/10_900x600_24054.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/11_900_x_600_24056.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/12_900x600_24058.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/13_900x600_24060.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/16_900x600_24062.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/17_900_x_600_24064.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/18_900x600_24066.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/19_900x600_24068.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/20_900x600_24070.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/2_900x600_1_24072.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/04_900_x_600_1_24076.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/3_900x600_1_24074.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/01_1_24841.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/02_4_24861.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/03_1_24863.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/04_1_24865.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/05_1_24867.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/06_24869.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/07_1_24877.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/08_1_24879.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/09_1_24881.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/10_24883.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/11_24885.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/13_24889.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/14_1_24899.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/15_1_24901.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/16_1_24903.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/17_1_24905.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/19_24909.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/20_24911.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/18_2_24925.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/12_3_24929.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/07_done_28292.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/08_done_1_28308.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/06_done_2_28294.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/11_done_1_28310.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/13_done_2_28328.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/12_done_1_28312.jpg"},{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/done_28428.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/output_img_0824_fong_gjh_28408.jpg"},
{"src":"http:\/\/\/\/d1kjvfsq8j7onh.cloudfront.net\/img\/output_img_4727_lauchingfong_amanjaya_fr_28432.jpg"}]}
```

Figure 5. JSON format

The experimentation for deep web, the web data extraction is performed by considering the size and different level of images [27]. This experiment has been conducted with regards to former works done by [16] to compare the performances of extraction process. The image extraction has been extracted in three ways:

a) The extraction of images in general way
b) The extraction of images by considering the size of images in two parts; 50*50 pixels and 128*128 pixels.
c) The extraction of images is tested randomly at different levels; 5 pages, 10 pages, 15 pages, 20 pages, 25 pages and 30 pages.

In this paper, we discuss the result of deep web data extraction by extraction of images that has been tested randomly for 30 pages as shown in Tables 2 and 3 and by considering two parts of extraction pixels; 50*50 pixels and 128*128 pixels. Table 4 shows the percentage of time extraction regarding to Table 4(a) and (b). From this table, we can see that the percentage of time extraction for WEIDJ and WEIDJ-no rules is lower compared to image extraction using DOM and WHDJ. This performance can prove that the extraction semi-structured data using WEIDJ is fastest compared to others.

Table 2. Performance of image extraction by web pages (30 URL) for DOM and WHDJ

Benchmark	DOM				WHDJ			
	Image found	Image retrieved	Image filtered	Time	Image found	Image retrieved	Image filtered	Time
amnh.org	1662	611	1051	3845.7278	1077	578	499	2457.5042
ocean.si.edu	687	610			77	751.5967	62	715.2595
iucn.org	289	251	38	683.3783	227	191	36	509.2624
endangeredspeciesinternational.org	77	43	34	158.5747	59	43	16	116.4149
wwf.org.my	492	375	117	503.206	460	371	89	462.0894

Table 3. Performance of image extraction by web pages (30 URL) for WEIDJ and WEIDJ (no-rules)

Benchmark	WEIDJ				WEIDJ(no-rules)	
	Image found	Image retrieved	Image filtered	Time	Image retrieved	Time
amnh.org	249	204	45	100.272	5430/1691	510.6992
ocean.si.edu	379	366	13	82.7162	691/676	254.8985
iucn.org	118	101	17	108.7956	819/274	208.7372
endangeredspeciesinternational.org	277	105	172	47.4335	427/401	38.9521
wwf.org.my	371	276	94	94.9288	495/461	77.9276

Table 4. Performance of Image extraction by percentage for 30 URL

Web address	DOM		WHDJ		WEIDJ		WEIDJ-no rules	
	Time	Percentage %	Time	Percentage %	Time	Percentage %	Time	Percentage %
amnh.org	3845.7278	55.6	2457.5042	35.5	100.272	1.5	510.6992	7.4
ocean.si.edu	751.5967	42	715.2595	40	82.7162	4	254.8985	14
iucn.org	683.3783	45	509.2624	34	108.7956	7.2	208.7372	13.8
endangeredspeciesinternational.org	158.5747	43.9	116.4149	32.2	47.4335	13.1	38.9521	10.8
wwf.org.my	503.206	44.2	462.0894	40.6	94.9288	8.35	77.9276	6.85

To give better visualization for users, Figure 6 shows the performance of time for each model in extracting images for WWF website (refer to Table 4). From this figure, we can see that time performance of Document Object Model is 44% which is contributing longer than other models. This is because the model needs to check the images for each node one by one before extracting all images from this website. The wrapper hybrid DOM and JSON (WHDJ) has been proposed to overcome the limitation of DOM. The results show the hybrid model, combination of DOM and JSON (40%) is success. However, although the time has been reduced but there are certain images that can not been extracted. That is the weakness of WHDJ. So, in this research work we proposed a new hybrid model which is combination of the visual segmentation and handling noisy images can be detected to ensure that only beneficial images can be retrieved. The definition of noisy images is the images that may contains of logo, repetition of images and many more. This is because web, despite acts as large repositories of knowledge, it undeniably also contains noisy information. Noisy information can degrade the performances of data extractions. WEIDJ is proposed in order to overcome the limitation of extracting beneficial images and remove noisy images to ensure it can extract images in fastest way. From this figure, the percentage of WEIDJ in extracting images is quite fastest (8%). WEIDJ No-rules is implementing similar technique in WEID model but this model will retrieve all types ofimages inclusing noisy images.

Table 5 and 6 shows image extraction for deep web that have sample size of image between 50x50 and 128x128. The reason the extraction has been experimented in between this two size is because the beneficial image size normally in rage 128x128 but the noisy images such as header, logo and so forth is in 50x50 pixels.

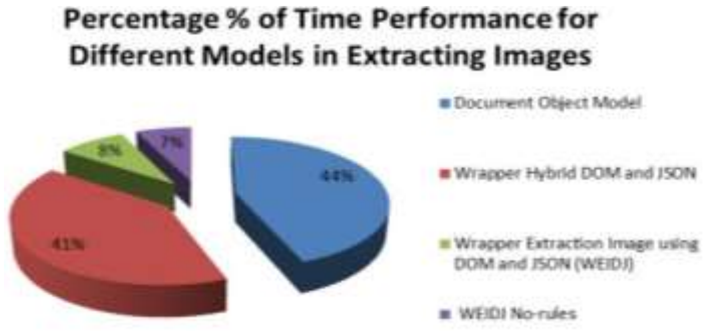


Figure 6. Percentage % of time performance for different models in extracting images

Table 5. Performance of image extraction for deep web (Size 50*50)

Benchmark	Link Found	Img found	Img retrieved	Img filtered	Time
			DOM		
amnh.org	132	4881	2125	2756	10556.2238
ocean.si.edu	97	1966	1610	356	2319.4244
iucn.org	96	999	811	188	1979.6851
endangeredspeciesinternational.org	30	394	288	96	865.8827
wwf.org.my	142	1803	1374	429	1900.9394
			WHDJ		
amnh.org	132	4013	2028	1985	8778.3747
ocean.si.edu	97	1705	1505	200	2076.7548
iucn.org	96	707	596	111	1625.4596
endangeredspeciesinternational.org	30	300	269	31	534.6634
wwf.org.my	142	1626	1370	256	1385.7157
			WEIDJ		
amnh.org	132	1521	1385	136	457.7495
ocean.si.edu	96	836	803	33	312.985
iucn.org	96	340	310	30	308.6347
endangeredspeciesinternational.org	30	262	102	160	26.4048
wwf.org.my	143	1311	1059	251	318.2913
			WEIDJ (no Rules)		
amnh.org				7339/4921	928.7615
ocean.si.edu				3832/1972	580.42
iucn.org				1952/1011	660.984
endangeredspeciesinternational.org				427/401	36.8205
wwf.org.my				3672/1907	573.7713

Table 6. Performance of image extraction for deep web (Size 128*128)

Benchmark	Link Found	Img found	Img retrieved	Img filtered	Time
			DOM		
amnh.org	133	4920	839	4081	13709.2253
ocean.si.edu	97	2007	404	1603	3244.0467
iucn.org	96	998	493	505	2980.6396
endangeredspeciesinternational.org	30	394	78	316	808.1518
wwf.org.my	143	1818	307	1515	1621.6796
			WHDJ		
amnh.org	134	4124	822	3302	12223.65
ocean.si.edu	98	1681	404	1277	1888.9131
iucn.org	97	790	523	267	1772.138
endangeredspeciesinternational.org	30	300	66	234	436.362
wwf.org.my	143	1175	164	1011	592.1318
			WEIDJ		
amnh.org	1593	1420	173	1697.7931	1593
ocean.si.edu	98	846	807	39	1368.3641
iucn.org	97	389	330	59	1253.8517
endangeredspeciesinternational.org	30	277	93	184	45.6617
wwf.org.my	143	1371	541	829	342.2131
			WEIDJ(no-rules)		
amnh.org				7012/4918/	1335.5362
ocean.si.edu				3902/2005	533.1249
iucn.org				1002/970	540.0529
endangeredspeciesinternational.org				400/427	31.1268
wwf.org.my				2541/1346	310.7469

4. CONCLUSION

In this paper, we have described a model for web data extraction programs, which provides an offer potential web data extraction for users. Among 17 websites that we used for the evaluation experiment, the experimental work discusses the extraction from five websites and the level of extraction is focusing on deep web. It can be operated by the action of users in clicking and pointing the cursor to search the web address after inserting the web *url*. This experiment shows that our proposed wrapper is able to reduce user's burdern in writing any configuration file due to different structure of each web page although it is in the same website. An important part of our work was the model of web data extraction, the execution time of extraction become longer especially in extracting a large numbers of images due to contain the noisy images also. Majority of the techniques convert the website into DOM tree so that they can be analyzed to identify noises by removing the unrelated elements. The extraction time becomes longer so an alternative have been conducted to decrease the execution time by applying JSON in WHDJ. An improved algorithm and better solution in dealing with the ever expanding data size, which would further complicate the processing of the data, should be invented. After the extraction is successful the images and related information will be saved in a database as a structured format. This information can be used for further action such as decision making. The one relevant of this extraction process is the execution time is reduce and the image's filenames will be reindexed. In future work, we are planning to extend this research work in focusing extraction from multi deep websites. The performance of images extraction will influence the time for execution process and the impact of the study for the nation and community is the extraction of semi-structured data that can be used for managing and analyzing the characteristics of elements.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] S. Z. Z. Abidin, N. M. Idris, A. H. Husain, "Extraction and classification of unstructured data in WebPages for structured multimedia database via XML," *International Conference on Information Retrieval & Knowledge Management (CAMP)*, 2010, pp. 44-49, doi: 10.1109/INFRKM.2010.5466948.

[2] D. Cai, S. Yu, J. Wen, W. Ma, "VIPS: A Vision-Based Page Segmentation Algorithm," *Book VIPS: a vision-based page segmentation algorithm*, Microsoft technical report, MSR-TR-2003-79, 2003.

[3] Z. Cai, J. Liu, L. Xu, C. Yin, J. Wang, "A Vision Recognition Based Method for Web Data Extraction," *Computer Science*, 2017.

[4] Chia-Hui Chang, Shih-Chien Kuo, "Olera: semisupervised Web-data extraction with visual support," *IEEE Intelligent Systems*, vol. 19, no. 6, pp. 56-64, Nov.-Dec. 2004, doi: 10.1109/MIS.2004.71.

[5] Chia-Hui Chang, Shao-Chen Lui, "IEPAD: Information Extraction Based on Pattern Discovery," *Book IEPAD: Information extraction based on pattern discovery'ACM*, pp. 681-688, 2001.

[6] M. Citra, A. A. Banu, "Deep Web Data Extraction Based on URL and Domain Classification," *ISAACA Journal*, vol. 4, pp. 1-4, 2015,

[7] V. Crescenzi, G. Mecca, P. Merialdo, "Roadrunner: Towards Automatic Data Extraction from Large Web Sites," *Book Roadrunner: Towards automatic data extraction from large web sites*, pp. 109-118, 2001.

[8] N. Derouiche, B. Cautis, T. Abdessalem, "Automatic Extraction of Structured Web Data with Domain Knowledge," *IEEE 28th International Conference on Data Engineering*, 2012, pp. 726-737, doi: 10.1109/ICDE.2012.90.

[9] Y. Fang, X. Xie, X. Zhang, R. Cheng, Z, Zhang, "STEM: A Suffix Tree-Based Method for Web Data Records Extraction," *Knowledge and Information Systems*, vol. 55, no. 2, pp. 305-331, 2018.

[10] P. Gulati, M. Yadav, "A Novel Approach for Extracting Pertinent Keywords for Web Image Annotation using Semantic Distance and Euclidean Distance," *Software Engineering*, pp. 173-183, 2019.

[11] D. T. Hai, "A Novel Integer Linear Programming Formulation for Designing Transparent WDM Optical Core Networks," *International Conference on Advanced Technologies for Communications ATC*, 2019, pp. 273-277, doi: 10.1109/ATC.2019.8924515.

[12] J. Hammer, G. Molina, H. Cho, R. Aranha, A. Crespo, "A.: 'Extracting Semistructured Information from the Web," Standford Infolab Publication Server, 1997.

[13] C. N. Hsu, M. T. Dung, "Generating Finite-State Transducers for Semi-Structured Data Extraction from The Web," *Information Systems*, vol. 23, no. 8, pp. 521-538, 1998, doi: https://doi.org/10.1016/S0306-4379(98)00027-1.

[14] R.Jefferson, A.Connell, and O. Jefferson, "Web Data Extraction", Lens.org, "Web Data Extraction", accessed 21 April 2021.

[15] P. Jimenez, R. Corchuelo, "On Learning Web Information Extraction Rules with TANGO," *Information Systems*, vol. 62, pp. 74-103, 2016, doi: https://doi.org/10.1016/j.is.2016.05.003.

[16] N. V. Kamanwar, S. G. Kale, “Web data extraction techniques: A review,” *World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave)*, 2016, pp. 1-5, doi: 10.1109/STARTUP.2016.7583910.

[17] A. H. F. Laender, B. A. R. Neto, A. S. Da silva, J. S. Teixeira, “A Brief Survey of Web Data Extraction Tools,” *ACM Sigmod Record*, vol. 31,no. 2, pp. 84-93, doi: https://doi.org/10.1145/565117.565137.

[18] B. Liu, R. Grossman, Y. Zhai, “Mining data records in web pages,” *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 601-606, 2003.

[19] L. Liu, C. Pu, W. Han, “XWRAP: an XML-enabled wrapper construction system for Web information sources,” *Proceedings of 16th International Conference on Data Engineering* (Cat. No.00CB37073), 2000, pp. 611-621, doi: 10.1109/ICDE.2000.839475.

[20] P. Malhotra, S. K. Malik, “Web Page Segmentation Towards Information Extraction for Web Semantics,” *International Conference on Innovative Computing and Communications*, pp. 431-442, 2018.

[21] M. Man, I. A. A. Sabri, M. M. A. Jalil, N. Ali, S. Muhamad, “Information Integration Architecture System for Empowering Rural Woman In Setiu Wetlands, Terengganu, Malaysia,” *Journal of Sustainability Science and Management*, vol, 14, no. 1, pp. 77-86, 2019.

[22] A. Pouramini, S. K. Hassani, Sh. Nasiri, “Data Extraction Using Content Based Handles,” *Journal of AI and Data Mining*, vol. 6, no. 2, pp. 399-407, 2018, doi: 10.22044/JADM.2017.990.

[23] I. A. A. Sabri, M. Man, “A Performance of Comparative Study for Semi-Structured Web Data Extraction Model,” *International Journal of Electrical and Computer Engineering*, vol. 9, no. 6, pp. 5463-5470, 2019, doi: 10.11591/ijece.v9i6.pp5463-5470.

[24] I. A. A. Sabri, M. Man, “Improving Performance of DOM in Semi-Structured Data Extraction Using WEIDJ Model,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 9, no. 3, pp. 752-763, 2018, doi: 10.11591/ijeecs.v9.i3.pp752-763.

[25] I. A. A. Sabri, M. Man, “WEIDJ: Development Of A New Algorithm For Semi-Structured Web Data Extraction,” *TELKOMNIKA* (*Telecommunication Computing Electronics and Control)*, vol. 19, no. 1, pp. 317-326, 2021, doi: 10.12928/TELKOMNIKA.v19i1.16205.

[26] A. K. Tripathy, N. Joshi, S. Thomas, S. Shetty and N. Thomas, “VEDD- a visual wrapper for extraction of data using DOM tree,” *International Conference on Communication, Information & Computing Technology ICCICT*, 2012, pp. 1-6, doi: 10.1109/ICCICT.2012.6398114.

[27] I. A. A. Sabri, M. Man, “Performance Analysis for Mining Images of Deep Web,” *International Journal of Advanced Computer Science and Applications IJACS*, 2020, vol. 11, no. 10, pp. 1-7, 2020, doi: 10.14569/IJACSA.2020.0111001.

BIOGRAPHIES OF AUTHORS

Ily Amalina Ahmad Sabri, received her Diploma in Information Technology in 2006 from PSMZA, Terengganu, Bachelor of Information Technology (Software Engineering), Master's degree, and Ph.D. in Computer Science from Universiti Malaysia Terengganu in 2009, 2014, and 2019 respectively. She is a Senior Lecturer in Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu. Her research interests include Web Mining, Data Extraction, Information Retrieval, Artificial Intelligence and Decision Support System. Her current research projects are "M-FlyCounter Development of Auto-Counting Mobile Apps for Large Populations of Housefly for Pest Control and Monitoring Activities” which is funded by PPRG 2021 scheme, Kajian dan Pembangunan Perisian untuk Studio Al-Quran, UMT which is funded by UMT, iMAKERS@UMT which is funded by MOSTI, “Development and Implementation of DIETCARE: An Interactive Online Nutritional Database Management System to support Intelligent Client Monitoring” which is funded by TAPE-RG and “An intelligent Tissue Dispenser System” which is funded by PPRG.

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, Master 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.