ScienceDirect Web Data Extraction Approach for Deep Web using WEIDJ

Conference Paper · July 2019

CITATIONS
0

READS
354

4 authors, including:

Ily amalina ahmad sabri
Universiti Malaysia Terengganu
14 PUBLICATIONS **23** CITATIONS
SEE PROFILE

Mustafa Man
Universiti Malaysia Terengganu
74 PUBLICATIONS **168** CITATIONS
SEE PROFILE

Wan Aezwani Bt Wan Abu Bakar
Universiti Malaysia Terengganu
19 PUBLICATIONS **13** CITATIONS
SEE PROFILE

Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 00 (2019) 000–000

Procedia Computer Science
www.elsevier.com/locate/procedia

16th International Learning & Technology Conference 2019

Web Data Extraction Approach for Deep Web using WEIDJ

Ily Amalina Ahmad Sabri[a]*, Mustafa Man[a]†, Wan Aezwani Wan Abu Bakar[b], Ahmad Nazari Mohd Rose[b]

[a]*School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu Terengganu, Malaysia*
[b]*Faculty of Informatics and Computing, University Sultan Zainal Abidin, Terengganu, Malaysia*

Abstract

Data extraction is one of the most prominent areas in data mining analysis that is been extensively studied especially in the field of data requirements and reservoir. The main aim of data extraction with regards to semi-structured data is to retrieve beneficial information from the World Wide Web. The data from large web data also known as deep web is retrievable but it requires request through form submission because it cannot be performed by any search engines. Data mining applications and automatic data extraction are very cumbersome due to the diverse structure of web pages. Most of the previous data extraction techniques were dealing with various data types such as text, audio, video and etc. but research works that are focusing on image as data are still lacking. Document Object Model (DOM) is an example of the state of the art of data extraction technique that is related to research work in mining image data. DOM was the method used to solve semi-structured data extraction from web. However, as the HTML documents start to grow larger, it has been found that the process of data extraction has been plagued with lengthy processing time and noisy information. In this research work, we propose an improved model namely Wrapper Extraction of Image using DOM and JSON (WEIDJ) in response to the promising results 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. To observe the efficiency of the proposed model, we compare the performance of data extraction by different level of page extraction with existing methods such as VIBS, MDR, DEPTA and VIDE. It has yielded the best results in Precision with 100, Recall with 97.93103 and F-measure with 98.9547.

Keywords: Document Object Model; Web Data Extraction; Wrapper Extraction of Image using DOM and JSON (WEIDJ)

* Corresponding author e-*mail address:* ilylina@yahoo.com
* Corresponding author e-*mail address:* mustafaman@umt.edu.my

1. Introduction

Data mining is the process of extracting the useful and relevant information from large amount of database. The extraction of the information from the large database is called as Knowledge Discovery (KD). It allows user to analyze data from different views and then categorize 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 current trend is that the numbers of devices and gadgets connection to the Internet is on the rise. This increase in connections to the internet has created the web as the largest source of information worldwide. With the large amount of data residing in the web, and complemented by advanced technologies in database processing, it is therefore a seamless effort to gather, collect and process the data. As the consequence of the exponential data growth, it is most important for users to adopt advanced data analytics technologies to efficiently store, retrieve and analyze the data. The main aim is to utilize this data, to learn about patterns and trends that can be used to make a positive impact on our lifestyle. However, the data itself doesn't produce these objectives, but rather its solutions that arise from analyzing it and finding the answers we need. This accumulation of data in terms of volume, technology and techniques are often being discussed in relation to big data, knowledge discovery in database, data mining, data integration and data extraction.

1.1. Big Data

Big data is a collection of huge amount of data that requires special database management systems to retrieve and analyse useful insight from it [1]. The term big data encompasses a wide range of approaches of collecting and analysing data in ways that were not possible before the era of modern personal computing. One approach to big data of great potential to end users is web scraping, which involves the automated collection of information from web pages [2]. The analytics from big data will enable us to find new cures and to better understand about some disease and healthcare [3, 4]. The main reason why big data has become so important is that its application has found its way into most of the fields and the masses are benefitting positively from it [5]. It represents a new era in data exploration. It is a large data set that need tools or applications for data processing purpose [6, 7]. The data obtained is not volatile of nature and thus can be easily processed and understood with the help from computers and mathematical models. The technologies of big data are particularly adapt at handling all types of data; such as unstructured data and semi-structured data. These data has grown explosively and requires new techniques and tools that can transform the processed data into useful knowledge automatically[8]. One of the processes used to transform the processed data into knowledge is Knowledge Discovery Database (KDD).

1.2. Knowledge Discovery in Database (KDD)

Knowledge Discovery in Database (KDD) is an interdisciplinary area focusing on methodologies or techniques for extracting useful data from sources [9]. The challenge of extracting useful data has drawn the interest of researchers to further investigate the data extraction in the field of databases, pattern recognition, machine learning, data visualization and high performance computing. The process in KDD by assigning a set of intelligent agents in each phase improves communication and cooperation between the different KDD steps in order to generate relevant results knowledge for decision-making tasks and deliver advances intelligence and web discovery solutions including using web mining techniques [10] [11]. Lots of researches have been accomplished in data integration and data extraction. Fig. 1.1 shows simple analogy for Knowledge Discovery Process (KDD) in extracting knowledge to users. It shows that there are five main steps for KDD process; Selection, Pre-processing, Transformation, Data Mining and Evaluation. During selection step, user must be clear and set the main goal of collecting data such as collect data from variety of sources (data warehouse) and then integrate the data into a single data stored. Then, relevant data or known as target data to the analysis task will be retrieved. During the second phase, pre-processing,

target data will be pre-processed. Data such as noise information and inconsistent data will be removed. In addition, data that comes from multiple sources may be combined such as flat files, spread sheets and relational database. In transformation process, data are transformed or consolidated into forms appropriate for mining purpose by performing summary or aggregation operations. Later, in data mining phase, data mining algorithms acts as pattern and rules that will be used to process the data to become output. It is an essential process such as clustering, classification and regression where data mining methods (algorithms) are applied in order to extract data. Finally, evaluation as the last phase, is very important step to identify the patterns which is representing knowledge based on certain measurement.

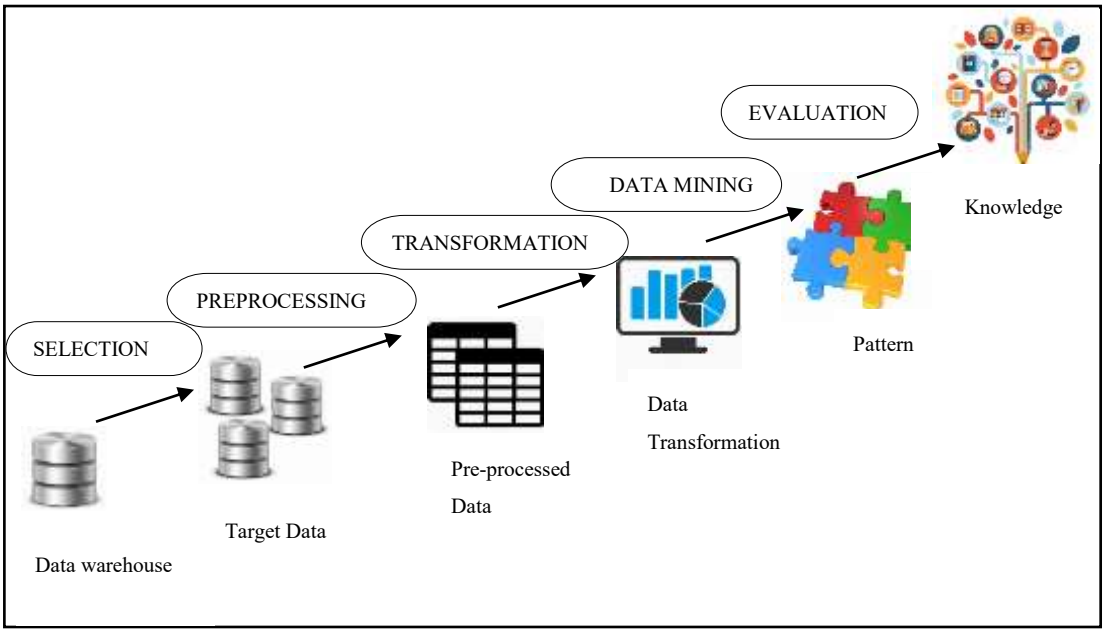

Fig. 1. Knowledge Discovery Database (KDD) Process

1.3. Data Mining

Data mining uncover new facts and relationship using useful patterns and techniques in order to give a solution for handling big data [12]. Data mining techniques are implemented to find useful patterns in large database such as MySQL and Oracle. It is the process that tries to discover patterns or techniques that can be applied in large dataset. The main goal of data mining is to extract information from large dataset. Enough data and supported tools are important and need to be completed each other in dealing with large data set. It may be leveraging onto the implementation of the big data that provides great opportunities for several of fields such as e-commerce, industrial control and smart medical [13].

1.4. Data Integration

Data integration involves combining data residing from different sources and providing users with a unified view of them [14]. The data sources are integrated to form a single virtual database. If the data sources conformed to different data models, then these data need to be transformed in to a common data model as a part of the integration process. Merging information from multi-sources is a process that becomes significant in a variety of situations in order to provide users the illusions of interacting in single view. There are two reasons why data need to be integrated. The first reason is data integration is important to user because the integration of data can provide users

wide viewpoint of data. The second reason is the combination of data from different sources can gain more comprehensive level of data that can fulfil user's requirement.

1.5. Data Extraction

Data extraction is where data is been analysed and crawled through from data sources such as web or database. It depends on specific patterns of user requirements. The goal of data extraction is to retrieve relevant information. It organizes data into usable and valuable resource so that we can use for further purpose. The extraction process may involves different data types.

The advantages of data extraction from semi-structured data is that it can be applied in various fields such as in education [15], advertisements [16], housing management [17]. In former work, data extraction that has been discussed can be modelled by using single model or combining several models together for the best assessment [18]. While web has developed into a large source of information, there are different data types of information that will be discussed in next section.

2. Noisy Information

Web pages consist of a lot of information which may include some information that can degrade the performance of extracting information. There has been lots of discussion with regards to noisy information. Liu et. al [19] and Yi and Liu [20] proposed taxonomy that consists of two categories of web noise include local noise and global noise.

i. Local Noise

Local noise is described as noises within a web page. It is also called as intra-web noise. The examples of local noise are unnecessary images, advertisements, privacy notice, links for navigation, copyright notice and so forth. Each web page contains a lot of information including the meaningful information and irrelevant information.

ii. Global Noise

Noise information within a website also known as global noise or inter-web page noise. The examples of global noise are such as unnecessary information, advertisement, links of navigation, copyright notice that can be found in a website.

iii. Prominent Global Noise

There is an opinion that noise information can be classified as prominent global. The noises may consist of mirrored website, duplicate web pages, the previous version of website and etc.[21].

iv. Navigational Noise

Besides that, web page noises can be summarised into three categories; fixed noise, web service noise and navigational noise.

- Fixed noises consist three types of noisy information; page description noise, decoration noise and statement noise.
- Decoration noise are such as logos, graphics, decorative text etc.
- Copyright details, terms and conditions, privacy statements etc. are examples of statement noises information.
- Date, time, visitor count etc. are different types of noisy information for page description noises.
- In addition, web pages also contain service blocks in the main information. These service blocks are important in performing certain task smoothly, communicating with the server and so forth.

3. Wrapper Extraction of Image Using DOM and JSON (WEIDJ) Model

WEIDJ is developed to assist user in extracting semi-structured data from web page. A web page can be represented by a tree structure Document Object Model (DOM). It converts and store a given web address of web page from a search engine into a DOM tree [22]. Recently, the extraction process is focused on image [23, 24]. When user input the uniform resource locator (URL) and the query is submitted to a search engine, the search engine will dynamically generated result page containing the result records. The results consists a link path for each element of image, image, size of image and time processing to load each image [25].

WEIDJ used AJAX technology to extract data from web sources. AJAX, is the abbreviation of Asynchronous JavaScript and XML, is a set of web development techniques that allows a web page to update portions of contents without having to refresh the page. AJAX represents a similar concept to the client-server development. During client-server phase, the amount of data transferred is very minimal over a terminal application by transferring only the necessary data back and forth. Similarly, with AJAX, only the necessary data is transferred back and forth between the client and the web server. This minimizes the network utilization and processing on the client. The time for extraction process has been reduced.

It can be difficult to properly create extraction rules describing required data. In this thesis, we propose WEIDJ [26] model to extract images from a web page as shown in Figure 2. The work described in this section uses a combination of both techniques, DOM and JSON [27]. In addition, we also do the checking of images by blocks in the HTML documents. It also focusses on arranging the extracted data in a tabular format. Lots of applications focuses on extracting information and then have it arranged accordingly [28, 29]. Every web page has their own structure includes main topic, related topics, additional information, advertisement, contact information, images, audio and video file. The problem that we want to solve is what is the best technique can be applied in order to extract images automatically[30]. Mining information records in data regions plays important role in defining tags of semi-structured data. It is easy to extract data from data regions because it contains useful data. It is recognized as data area. A technique is requisite in order to mine data area. In the earlier stage, this model proposed DOM tree as based technique to mine data regions in web page as discussed in previous chapter.

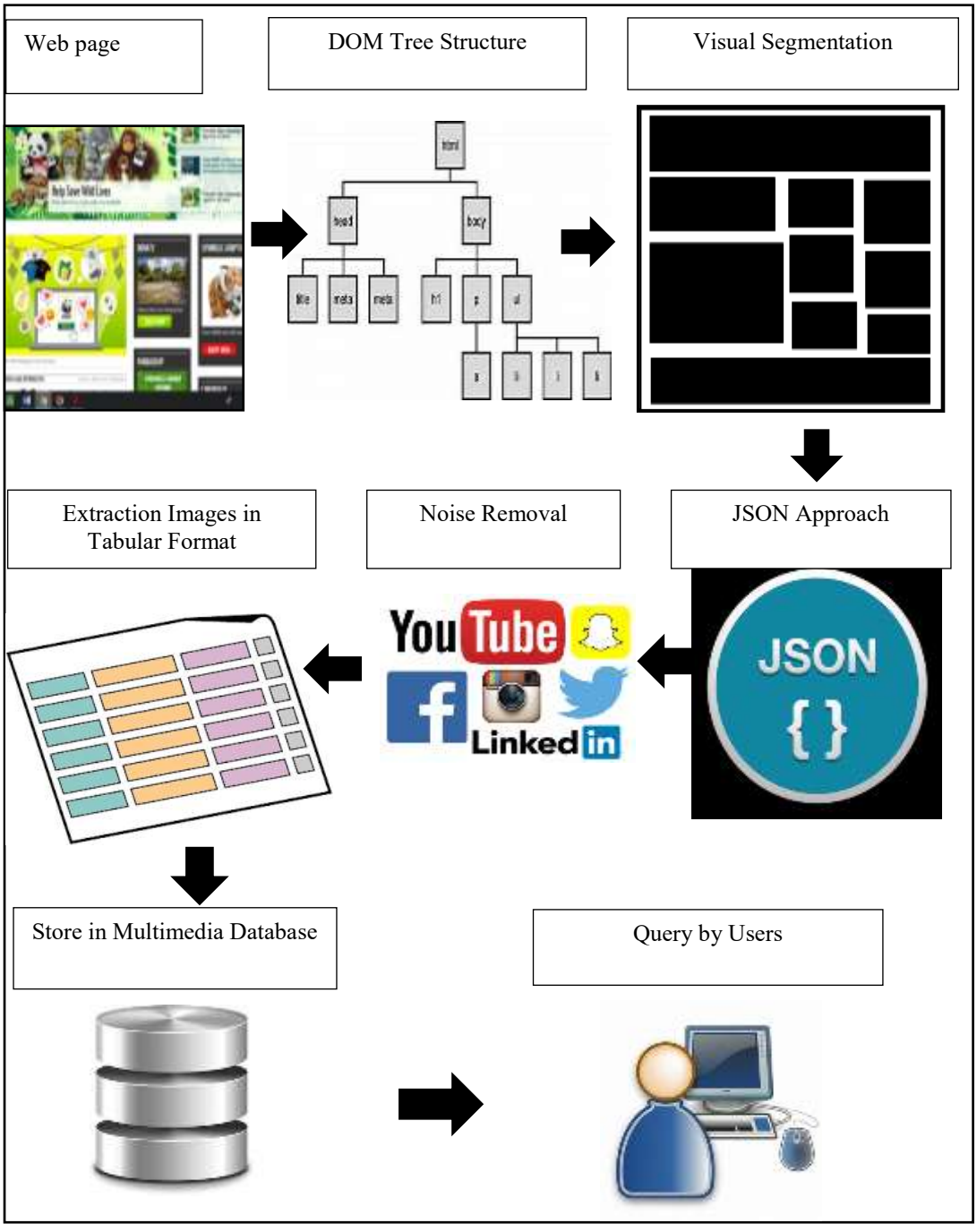

Fig. 2. Workflow for Extraction Images in WEIDJ

4. Comparison of Performance with Existing Methods

In this research work, web data extraction experiments also been set up to compare the performance of WEIDJ with existing method. We select the website of FangJia which is http://sh.FangJia.com. The reason why this website is selected as a guideline because there is a discussion of findings for image extraction that has been constructed [23]. Four typical data extraction algorithm VIBS, MDR, DEPTA and VIDE were selected as comparing target. The experiments were conducted on the prototype system of the above algorithm. There are two types of performance measurement that have been conducted during this experiment. The first measurement is execution time analysis and second is precision, recall and F-measure.

All experiments are performed on HP-UNOJQE6, Intel(R) Core(TM) i7-6500U CPU @250 GHz with 12.0 GB RAM in a Win10 64bit operating system platform. The software specification for algorithm development is deployed using open source software i.e MySQL version Apache/2.4.10 (Win32) OpenSSL/1.0.1i PHP/5.5.19 for our web server, php and JSON as programming language and phpMyAdmin with version 4.7.3, the latest stable version as to handle the administration of MySQL over the web. The phpMyAdmin (phpMyAdmin: Bringing MySQL to the web) is a free software tool written in PHP that supports a wide range of operations on MySQL. Frequently used operations (managing databases, tables, columns, indexes, etc) can be performed via the user interface. The software configuration is shown in Table 1 below.

Table 1. Software configuration

Software	Configuration
Database Server	• MySQL
Web Server	• Apache/2.4.10 (Win32) OpenSSL/1.0.1i PHP/5.5.19 • Database client version: libmysql - mysqlnd 5.0.11-dev - 20120503 - $Id: bf9ad53b11c9a57efdb1057292d73b928b8c5c77 $ • PHP extension: mysqli Documentation
MySQL Administration	• phpMyAdmin • Version information: 4.7.3, latest stable version: 4.7.3
Programming Language	• php • JSON

4.1. Time Extraction Analysis

In this experimental work, 40 pages from the same website (FangJia) has been selected randomly. Then, the extraction time will be calculated from the beginning of the extracted page to the next page. Fig. 3 shows output for extracting 5 pages for data extraction by corresponding page. The duration of the extraction process is shown in details in Table 2. From the performance analysis, we found that WEIDJ clearly outperforms compared to existing tools.

Table 2. The performance of data extraction

Method	Time Extraction							
	5 pages	10 pages	15 pages	20 pages	25 pages	30 pages	35 pages	40 pages
WEIDJ	**12.6972**	**18.639**	**22.18**	**29.1468**	**29.5079**	**35.2651**	**37.977**	**48.8498**
VIBS	7.25	12.7	23.74	30	35.01	44.37	49.76	62.69
MDR	19.29	40.11	61.18	83.78	101.07	122.63	148.33	164.16
DEPTA	20.98	43.79	66.66	90.63	114.04	135.72	153.55	180.71
VIDE	53.13	94.37	144.33	195.23	246.29	291.08	341.18	389.52

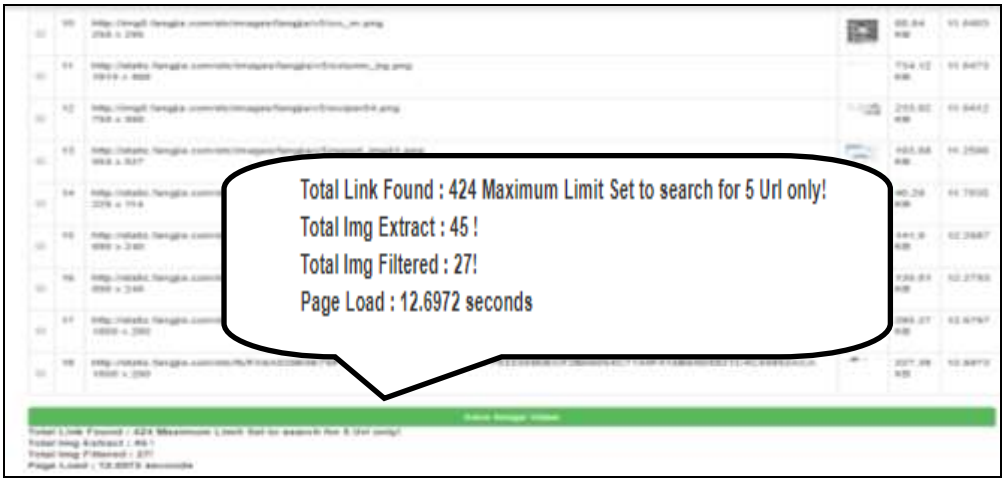

Fig. 3. Example of Extracting 5 Pages

4.2. Precision, Recall and F-measure

According to [23], the interference of web page noise to data extraction is important to be considered besides efficiency and accuracy of different deep web page heterogeneity. This issues motivates us to improvise existing algorithm on noisy information. So, besides focusing on the performance of time extraction for extracting information, we also want to extract the significant information of image and remove the noisy information. Table 3 shows the result of the experimental evaluation for WEIDJ using FangJia webpage as tested URL.

$$\Pr ecision = \frac{Dataretrieved}{Dataretrieved + Datafalse} \tag{1}$$

$$\mathrm{Re}\, call = \frac{Dataretrieved}{Totalofimage} \tag{2}$$

$$Fmeasure = 2\frac{\Pr ecision \bullet \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call} \tag{3}$$

Table 3. Result of the experimental evaluation for WEIDJ

Total img	Data retrieved	Data (False)	Unknown Data	Precision	Recall	F1
145	142	0	3	100	97.93103	98.9547

Fig. 4 shows the comparison of the five algorithm of the experiments. Our model, WEIDJ has proven that its ability to extract data as accurate as VIBS. This accuracy in extraction is achieved because of two factors that we include in this research, which are noises filteration and the use of JSON which helps to transform the data faster.

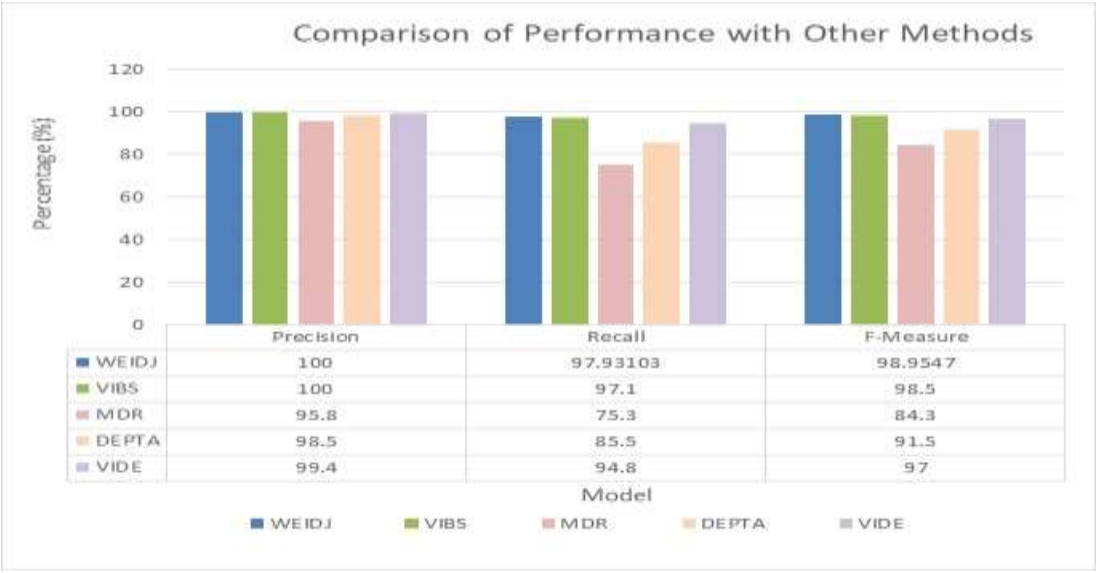

Fig. 4. Comparison Performance Existing Method

5. Conclusion

All the World Wide Web has become a vast information store that is growing at a rapid rate, either in number of sites or in volume of useful information. Web Data Extraction is time consuming when the html documents becomes larger. Single Document Object Model (DOM) did not perform very well in extracting multimedia data such as image if the volume of data become increased. However, when another technique JavaScript Object Notation is implemented in enhanced model namely as Wrapper Hybrid DOM and JSON (WHDJ), the time execution in extracting image and its information has been reduced to 50% greater than DOM technique. Even the time execution has improved but the limitation of this model is the redundancy of similar filename in images extraction. Complementary to this, we intend to combine both approaches and apply visual segmentation to get the best performance and extract the constructive images. This wrapper has been developed based on proposed model, Wrapper Extraction of Images using DOM and JSON (WEIDJ). The findings result of time execution of WEIDJ is greater (90%) than existing tools should be interpreted because of the three page level of extractions which are single URL, multiple URL and deep web used in the analysis of experimentation for the execution time.

In this study, the benchmark of dataset (18 websites) were heterogeneous with respect to image, path of images, size of images and execution time. Beside the execution time is focused as main guideline, the experimentation of images extraction would have improved the validity of significant information by removing noisy information of

images.

In future studies, it is recommended that the selection of dataset involves variety of fields which includes social networks or other platform. This is because the structure of website have been developed in different technologies.

References

1. Patel, B., et al., *Necessity of Big Data and Analytics for Good e-governance.* International Journal of Grid and Distributed Computing, 2017. **10**(8): 11-19.
2. Landers, R.N., et al., *A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data From the Internet for Use in Psychological Research.* Psychological Methods, 2016. **21**(4): 475-492.
3. Mehta, N. and A. Pandit, *Concurrence of big data analytics and healthcare: A systematic review.* International Journal of Medical Informatics, 2018. **114**: 57-65.
4. Wu, J., et al., *Decision based on big data research for non-small cell lung cancer in medical artificial system in developing country.* Computer Methods and Programs in Biomedicine, 2018. **159**: 87-101.
5. Ahmad, A., et al., *Multilevel Data Processing Using Parallel Algorithms for Analyzing Big Data in High-Performance Computing.* International Journal of Parallel Programming, 2018. **46**(3): 508-527.
6. Yang, W.Y., et al., *HEPart: A balanced hypergraph partitioning algorithm for big data applications.* Future Generation Computer Systems-the International Journal of Escience, 2018. **83**: 250-268.
7. Gai, K.K., et al., *In-memory big data analytics under space constraints using dynamic programming.* Future Generation Computer Systems-the International Journal of Escience, 2018. **83**: 219-227.
8. Man, M., et al., *Mining association rules: A case study on benchmark dense data.* Indonesian Journal of Electrical Engineering and Computer Science, 2016: 546-553.
9. De Oliveira, E.F., et al., *Voltage THD Analysis Using Knowledge Discovery in Databases With a Decision Tree Classifier.* Ieee Access, 2018. **6**: 1177-1188.
10. Ellouzi, H., M. ben Ayed, and H. Ltifi, *Modeling of Distributed visual Knowledge Discovery from Data Process*. 2017 12th International Conference on Intelligent Systems and Knowledge Engineering, ed. T. Li, L.M. Lopez, and Y. Li. 2017, New York: Ieee.
11. Lima, T.G., et al., *KDD Processes in Non-Relational Data: The case of the MineraMongo Tool*, in *2017 12th Iberian Conference on Information Systems and Technologies*. 2017, Ieee: New York.
12. Suresh, R., S.R. Harshni, and Ieee, *Data Mining and Text Mining - A Survey*, in *2017 International Conference on Computation of Power, Energy Information and Communication*. 2017, Ieee: New York. p. 412-419.
13. Zhang, Q.C., et al., *A survey on deep learning for big data.* Information Fusion, 2018. **42**: 146-157.
14. Gomez-Cabrero, D., et al., *Data integration in the era of omics: current and future challenges.* BMC systems biology, 2014. **8**(2): 1.
15. Williams, K., et al. *Scholarly big data information extraction and integration in the CiteSeer χ digital library*. in *Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on*. 2014. IEEE.
16. Pera, M.S., R. Qumsiyeh, and Y.-K. Ng, *Web-based closed-domain data extraction on online advertisements.* Information Systems, 2013. **38**(2): 183-197.
17. Dewaelheyns, V., I. Loris, and T. Steenberghen. *Web Data Extraction Systems versus Research Collaboration in Sustainable Planning for Housing: Smart Governance Takes It All*. in *REAL CORP 2016 Proceeding*. 2016.
18. 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 on*. 2016. IEEE.
19. Yi, L., B. Liu, and X. Li. *Eliminating noisy information in web pages for data mining*. in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003. ACM.
20. Yi, L. and B. Liu. *Web page cleaning for web mining through feature weighting*. in *IJCAI*. 2003.
21. Sivakumar, P. and R. Parvathi, *An efficient approach of noise removal from web page for effectual web content mining.* European Journal of Scientific Research, 2011. **50**(3): 340-351.
22. Sabri, I.A.A. and M. Man, *Multiple types of semi-structured data extraction using wrapper for extraction of image using DOM (WEID)*, in *Regional Conference on Science, Technology and Social Sciences (RCSTSS 2016)*, N.A.Y.e.a. (Ed.), Editor. 2018, Springer Nature Singapore Pte Ltd.: Singapore. p. 67-76
23. Liu, J., et al., *Deep web data extraction based on visual information processing.* Journal of Ambient Intelligence and Humanized Computing, 2017: 1-11.
24. Bhardwaj, A. and V. Mangat, *An improvised algorithm for relevant content extraction from web pages.* Journal of Emerging Technologies in Web Intelligence, 2014. **6**(2): 226-230.
25. Man, M. and I.A.A. Sabri, *The proposed algorithm for semi-structured data integration: Case study of Setiu wetland data set.* Journal of Telecommunication Electronic and Computer Engineering, 2017. **9**(No 3-3): 79-84.
26. Sabri, I.A.A. and M. Man. *WEIDJ : An improvised algorithm for image extraction from web pages*. in *The 8th international conference on information technology*. 2017. Al-Zaytoonah University of Jordan (ZUJ), Amman, Jordan: IEEE Xplore.
27. 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. **9**(3): 752-763.
28. Wang, J. and F.H. Lochovsky. *Data extraction and label assignment for web databases*. in *Proceedings of the 12th international conference on World Wide Web*. 2003. ACM.
29. Zhai, Y. and B. Liu. *Web data extraction based on partial tree alignment*. in *Proceedings of the 14th international conference on World Wide Web*. 2005. ACM.
30. Sabri, I.A.A. 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, 2017. **9**(3-3): 79-84.