A survey on technique for solving web page classification problem

IOP Conf. Series: Materials Science and Engineering **769** (2020) 012036 doi:10.1088/1757-899X/769/1/012036

A survey on technique for solving web page classification problem

Siti Hawa Apandi, Jamaludin Sallim and Rozlina Mohamed

Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang, Gambang, Kuantan, Pahang, Malaysia

E-mail: sitihawa.apandi@gmail.com

Abstract. Nowadays, the number of web pages on the World Wide Web has been increasing due to the popularity of the Internet usage. The web page classification is needed in order to organize the increasing number of web pages. There are many web page classification techniques that have been proposed by the other researchers. However, there is no comprehensive survey on the performance of the techniques for the web page classification. In this paper, surveys of the different web page classification techniques with the result of the techniques achieved are presented. The existing works of web page classification are reviewed. Based on the survey, we found that the neural network technique namely Convolutional Neural Network (CNN) produce high F-measure value and meet the real-time requirement for classification compared to the other machine learning technique.

1. Introduction

Many people have uses Internet especially World Wide Web as a platform to search information. The Web rapidly flourished with the increasing number of web pages. A web page is a a web document that contains information in forms of textual content, images, audio and video [1]. The web browser is needed in order for the Internet users to view the web pages. The common web browsers used are Google Chrome and Mozilla Firefox.

The increasing number of web pages has caused some problems which it is difficult for the user to get relevant result when searching information in the search engine. Besides that, it is hard to maintained indices up to date with the increasing number of web pages in the web search systems. Thus to solve this problems, the web page classification is needed. The accurate and efficient classification of the web pages can make the search engines to retrieve the required information by the users quickly and efficiently even though there are huge information on the web [2-5]. The classification of the web page is also important for the development of web directories, discussion of specific topics Web and contextual advertising links on the analysis of the structure current site [5, 6].

There are two methods for web page classification which are traditional manual method and automatic method [2]. The traditional manual method of the web page classification use human effort which is done by the expert to manually classify the web page. This manual method takes a great deal of human effort and time consuming. It is impossible to be used for the web page classification because there is increasing number of web pages nowadays [2]. Thus, the suitable method to be used

for the web page classification is the automatic method which uses the classifier in order to classify the web page.

There is an issue for the classifier which is huge scale dimensionality. This issue can be solved in the preprocessing phase of the web page classification process [2]. The preprocessing phase is very important process in the web page classification as it can provide better quality of training dataset. As a result, it can reduce the training and testing time. Besides, it can improve the performance result of the classifier [2, 7].

The other sections of this paper are structured as follows. In Section 2, the process of the web page classification is explained. This is followed by the review existing works of web page classification in Section 3. Lastly is the conclusion in Section 4.

2. Web page classification process

There is several process of web page classification as shown in Figure 1 which are collection of web pages, feature extraction, feature selection, build classifier, classify web pages and evaluate performance of classifier. The description of each process in web page classification is presented in the next subsection.

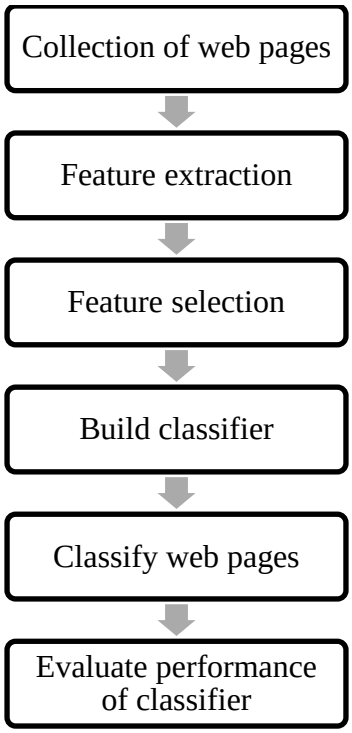

Figure 1. Process of web page classification.

2.1. Collection of web pages

Web pages are collected to be used as dataset in the web page classification. The dataset are split into two sets which are training and testing dataset. The training dataset is used to train the web page classifier, while the testing dataset is used to check the performance of the classifier [4].

2.2. Feature extraction

The feature extraction is one of the processes involved in the preprocessing phase that used to reduce the huge scale dimensionality issue. The irrelevant words and stop words that are found in the web pages is removed. The feature extraction process starts by extracting the raw content of the web pages with remove HTML tags and other WWW contents. Then, it continues with tokenization process

which means breaking down text into small chunks known as token that can be phrase, word or symbols. Stemming or lemmatization is done next to reduce the tokens to their root words. For example, the root word "move" is selected to represent the other similar words such as "moves", "moving" and "moved". Then, the stop words such as propositions, conjunctions and words that have high frequency are filtering out [2, 3].

2.3. Feature selection

The feature selection is another one of the processes involved in the preprocessing phase that used to reduce the huge scale dimensionality issue. This is important in order to increase the accuracy and efficiency of the classifier. The purpose of feature selection is to select the best features that would represent the web page [6]. The feature selection can be categorized into three methods which are filter, wrapper and embedded. Table 1 presents the comparison of the three feature selection methods.

Table 1. Comparison of feature selection methods.

Feature selection method	Filter	Wrapper	Embedded
Technique	Statistical measures	Optimization algorithm	Combination of filter and wrapper method
Computational efficiency	Efficient	Inefficient	Inefficient
Computation time	Time efficient	Slow	Slow
Computational cost	Cheaper	Expensive	Expensive
Computational space	Less computational space	More computational space	More computational space
Complexity	Low	High	High
Generality	High	Less	Less
For high dimensional data	Suitable	Not suitable	Not suitable
Advantage		High classification accuracy	Reduce the computation time
Disadvantage	Does not deal with redundant features	Increased runtime	

2.4. Build classifier

The selected features set are used as the input data set to be fed to the classifier. In training phase, the classifier is build by using machine learning algorithm.

2.5. Classify web pages

In testing phase, the classifier uses the learned function to classify the web page and allocate the web page to particular categories [2, 3].

2.6. Evaluate performance of classifier

The metrics used to evaluate the performance of classifier are precision, recall, F-measure and accuracy. The confusion matrix as shown in Table 2 is a table that presents the actual label of category again the predicted label of category [2]. The parameters in the Table 2 are used in order to compute the value of all the performance evaluation metrics which are precision, recall, F-measure and accuracy.

Table 2. Confusion matrix.

	Predicted positive	Predictive negative
Actual positive	True Positive (TP)	False Negative (FN)
Actual negative	False Positive (FP)	True Negative (TN)

The description of each the performance evaluation metrics for the web page classification are as follows.

- Precision tells us that based on the results classified as positive, how many were actually positive. The precision can be calculated by the formula as in (1).

$$\text{Precision (P)} = TP / (TP + FP) \quad (1)$$

- Recall tells us how many were actually correctly classified as True Positive based on the all positive data. It can be calculated by the formula as in (2).

$$\text{Recall (R)} = TP / (TP + FN) \quad (2)$$

- F-measure is calculated based on the value of precision and recall. The F-measure can be calculated by the formula as in (3).

$$\text{F-measure} = (2 \times P \times R) / (P + R) \quad (3)$$

- Accuracy tells us how many were correctly classified data based on the total number of data. The formula in (4) is used to calculate the accuracy.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

Of all the performance evaluation metrics discussed, it is better to use F-measure especially if there is an uneven class distribution which is the large number of actual negative. The accuracy is also useful to measure the performance but it is only when the values of False Positive and False Negative are almost same.

3. Existing works of web page classification

There are many techniques have been proposed by the other researchers for the web page classification. In this paper, we are going to review several existing works of web page classification. The Google Scholar is used to search for the papers related to the web page classification. The Google Scholar is used because there are many publication papers that can be found and it is easy to access. There are some criteria used to select the papers related to the web page classification that are going to be reviewed. Firstly, the paper published that is going to be reviewed need to discuss about the proposed technique to solve the web page classification. Then, the published paper selected is between the years of 2015 until 2018.

Table 3 presents a comparison of existing works of web page classification that has been arranged by descending order of year paper published. The comparison takes a look at the web page classification technique, the web page features used during the classification and the result achieved.

Table 3. Comparison of existing works of web page classification.

Author	Year	Title	Technique used	Feature of web page used	Performance evaluation metric used	Result
Shao, et al. [8]	2018	Effective Web-Page Classification Using Token-String CNN over URLs and Anchor Texts	• Token-string Convolutional Neural Network (CNN) based classifiers used to classify web pages	• URLs • Corresponding anchor texts	F-measure Classification time	Ranging from 0.93 to 0.99 Meets the real-time requirement. It take less than 3.2 seconds to classify 100K URLs
Wai, et al. [4]	2018	Ontology Based Web Page Classification System by Using Enhanced C4.5 and Naive Bayesian Classifiers	• The enhanced C4.5 decision tree and Naive Bayesian classifier used for classification of web pages • Use the ontology for semantic extraction about features	• Keyword • Semantic features	Accuracy	92.5%
Kiziloluk and Ozer [9]	2017	Web pages classification with Parliamentary Optimization Algorithm	• The Parliamentary Optimization Algorithm (POA) used to classify the web pages	• HTML tags	Accuracy	Ranging from 88.07% to 94.03%
Raj and Francis [10]	2017	Enhancements to Web Page Classification Based on Particle	• For feature selection, two optimization	• Terms and HTML tags	Precision Recall F-measures	92.21 87.69 89.89

		Swarm and Cuckoo Search Optimization	techniques namely Particle Swarm and Cuckoo Search used to refine and reduce feature sets • Naïve Bayes Classifier (NBC) used for classification of web pages		Accuracy	96.81
Bhatt, et al. [5]	2016	An Improved Optimized Web Page Classification using Firefly Algorithm with NB Classifier (WPCNB)	• The Firefly Algorithm (FA) used to optimize best feature selection • The Naive Bayes (NB) classifier is used for classification of web pages	• Term in HTML or XML tag	F-measure Accuracy	0.987 98.90%
Raj, et al. [11]	2016	Optimal web page classification technique based on informative content extraction and FA-NBC	• The Term Frequency -Inverse Document Frequency (TF-IDF) is used in the feature extraction phase • The Optimal Firefly Algorithm	• Terms and HTML tags	Precision Recall F-measure Accuracy	1.0 0.79 0.919 94.84

			(FA) based Naïve Bayes Classifier (FA-NBC) is used for classification of web pages			
Lee, et al. [3]	2015	Web page classification based on a Simplified Swarm Optimization	• The document frequency is used for feature selection • The Simplified Swarm Optimization (SSO) is used for web page classifier	• HTML tags and terms	Precision Recall F-measure CPU time	81.81% 83.46% 82.21% Require the greatest amount of the CPU time

Explanation about the feature selection methods as presented in Table 1 show that the filter method of feature selection is better than the other methods of feature selection. However based on the Table 3, we find that the technique involved in the existing works of web page classification mostly used optimization technique which is based on the wrapper method of feature selection [3, 5, 9-11]. This thing happens because the wrapper method of feature selection can produce high classification accuracy as stated in Table 1.

Based on the Table 3, we find that the commonly feature selected to be used in the existing works of web page classification is terms and HTML tags of web page [3, 10, 11].

From Table 3, we find the web page classifier that uses Convolutional Neural Network (CNN) which is one of the type of neural network, produce high value of F-measure [8]. This statement agreed by [7] that state the neural network work better than the other machine learning to be used as the web page classifier. It is also notice that the web page classification using CNN has advantage to meet the real-time requirement for the classification [8] compared to the work by Lee, et al. [3] which take greatest amount of CPU time.

4. Conclusion

This paper has discussed the basic idea of process and existing works in the web page classification. The process of web page classification includes collection of web pages, feature extraction, feature selection, build classifier, classify web pages and evaluate performance of classifier. The preprocessing process can be divided into two sub process which are feature extraction and feature selection. There are various techniques have been proposed for the web page classification. From the survey of the web page classification techniques, the Convolutional Neural Network (CNN) technique provides better result compared to other machine learning techniques. It achieves high F-measure value and meet real-time requirement for classification. This paper can be used as a guideline for the researchers who are interested to know the process and techniques involved in the web page

classification. For the future work, we would like to explore more on the neural network for the web page classification.

5. Acknowledgment

This research work is supported by Universiti Malaysia Pahang Grant: RDU182207-2. This work also partially supported by Adnuri SMA Research Center (M) Sdn. Bhd.

References

[1] E. Suganya and D. S. Vijayarani, "Web Page Classification in Web Mining Research - A Survey " *International Journal of Innovative Research in Science, Engineering and Technology,* vol. 6, pp. 17472-17479, 2017.

[2] A. Osanyin, O. Oladipupo, and I. Afolabi, "A Review on Web Page Classification," *Covenant Journal of Informatics and Communication Technology,* vol. 6, pp. 11-32, 2018.

[3] J.-H. Lee, W.-C. Yeh, and M.-C. Chuang, "Web page classification based on a simplified swarm optimization," *Applied Mathematics and Computation,* vol. 270, pp. 13-24, 2015.

[4] H. P. M. Wai, P. P. Tar, and P. Thwe, "Ontology Based Web Page Classification System by Using Enhanced C4. 5 and Naïve Bayesian Classifiers," in *2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, 2018, pp. 286-291.

[5] K. Bhatt, A. Singh, and D. Singh, "An Improved Optimized Web Page Classification using Firefly Algorithm with NB Classifier (WPCNB)," *International Journal of Computer Applications,* vol. 146, pp. 15-21, 2016.

[6] J. Alamelu Mangai, V. Santhosh Kumar, and V. Sugumaran, "Recent Research in Web Page Classification–A Review," *International Journal of Computer Engineering & Technology (IJCET),* vol. 1, pp. 112-122, 2010.

[7] L. Safae, B. El Habib, and T. Abderrahim, "A Review of Machine Learning Algorithms for Web Page Classification," in *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, 2018, pp. 220-226.

[8] L. Shao, S. Yao, X. Zhou, J. Guo, and J. Wang, "Effective Web-Page Classification Using Token-String CNN over URLs and Anchor Texts," in *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, 2018, pp. 105-111.

[9] S. Kiziloluk and A. B. Ozer, "Web pages classification with parliamentary optimization algorithm," *International Journal of Software Engineering and Knowledge Engineering,* vol. 27, pp. 499-513, 2017.

[10] A. M. J. Raj and F. S. Francis, "Enhancements to Web Page Classification Based on Particle Swarm and Cuckoo Search Optimization," *International Journal of Computer Science and Information Security (IJCSIS),* vol. 15, pp. 107-113, 2017.

[11] A. Raj, F. S. Francis, and P. J. Benadit, "Optimal web page classification technique based on informative content extraction and FA-NBC," *Computer Science and Engineering,* vol. 6, pp. 7-13, 2016.