2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)

2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM) | 978-1-6654-1407-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSECS52883.2021.00129

Data pre-processing of website browsing record: An initial step for web page classification

Siti Hawa Apandi
Faculty of Computing, College of Computing and Applied Sciences
Universiti Malaysia Pahang
Pekan, Pahang, Malaysia
sitihawa.apandi@gmail.com

Jamaludin Sallim
Faculty of Computing, College of Computing and Applied Sciences
Universiti Malaysia Pahang
Pekan, Pahang, Malaysia
jamal@ump.edu.my

Rozlina Mohamed
Faculty of Computing, College of Computing and Applied Sciences
Universiti Malaysia Pahang
Pekan, Pahang, Malaysia
rozlina@ump.edu.my

***Abstract*—The Internet utilization has resulted in an increase in the number of web pages on the World Wide Web. The classification of web pages is required to organize the growing number of web pages. A web page classification system is proposed to be constructed using a deep learning algorithm. The initial step for web page classification is data pre-processing. The website browsing record is used as a dataset in this study. The raw dataset needs to be pre-processing to fetch the cleaned data by removing missing value data, redundant data, and error data. There are many steps in data pre-processing which include data cleaning and web content pre-processing. The main contribution of this paper is to investigate how to do data pre-processing on website browsing records that focusing on the Game and Online Video web pages that will be utilized as the dataset to construct the web page classification model. After doing the data pre-processing, the number of datasets will be reduced. This shows many datasets have been removed because it is inactive and not suitable to be used in this study as the dataset of Game and Online Video web pages.**

Keywords— web page classification, data pre-processing, data cleaning, web content pre-processing, website browsing record

I. INTRODUCTION

Many people have utilized the internet, particularly the World Wide Web (often known as the Web), to conduct information searches. The Web grew quickly with a rising number of web pages. The exponential proliferation of web pages makes data extraction and categorization problematic for internet users. For example, when users search for particular information on a given topic, general purpose search engines frequently return too many irrelevant results. Even if there is a lot of information on the web, the classification of web pages can help search engines find the information that people need quickly and efficiently [1-4]. The classification of a web page is important not only for improving search engine performance, but also for the construction of web directories, discussion of specific web topics, and contextual advertising links on the study of the present site's structure [4, 5].

Internet usage has a positive and negative impact. The positive impact of internet usage makes it the people easy to search for information on the web. For the negative impact, people are exposed to internet addiction which refer to excessive internet use that disrupts daily living. Addiction to the internet causes problems for students in terms of academic performance, health, and social relationships [6].

The work presented in this paper is an extension of the paper entitled "A survey on technique for solving web page classification problem" by the authors published in IOP Conference Series: Materials Science and Engineering, 2020.

There is a confusion between Internet addicts and workaholic behaviors, as these people spending too much time in front of the computer and always stay connected to the internet. Souligna [7] states that the user that can be categorized as an Internet addict is regularly accessed the web page with the purpose of recreational Internet use rather than for business or school purposes. To prevent the students in the institution from getting internet addiction, the institution needs to set rules when the students using the internet provided by the institution and limit the usage of the internet towards recreational internet use.

The development of a system for web page classification utilizing a deep learning algorithm called Convolutional Neural Network (CNN) is proposed. We focus on topic classification for web page classification, which is classifying web pages based on their contents. We aim to find out if one web page is a Game or Online Video. These web page categories are selected because these are web pages regularly accessed by the users during their free time and it is related to recreational internet use. There are five steps required to develop the web page classification model which are data collection, data pre-processing, obtaining the required features, building the web page classification model, and evaluating the web page classification model as shown in Fig. 1.

978-1-6654-1407-4/21/$31.00 ©2021 IEEE
DOI 10.1109/ICSECS52883.2021.00129