IJARSCT

ISSN (Online) 2581-9429

IJARSCT
Impact Factor: 4.819

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 6, Issue 1, June 2021

Web Page Categorization using RNN Based on URL

Pratiksha Vaishnav[1] and Ankit Kalariya[2]
Computer Department, Atmiya University, Rajkot, India[1]
Computer Department, Atmiya University, Rajkot, India[2]

Abstract: *In this paper, Web page classification is an information retrieval application that provides useful information that can be a basis for many different application domains. 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. This research is going to simplify this problem through applying web mining techniques. Web page classification, also known as a web page categorization, to classify web pages categories based on URL.*

Keywords: Web Classification, Web Mining, Classification algorithms, RNN

I. Introduction

Web page classification is an information retrieval application that provides useful information that can be a basis for many different application domains. Categorizing web pages provides useful information for efficient internet use, spam filtering and many other application areas. Finding relevant results quickly from millions of web sites is a serious problem that must be solved for search engines. Web pages classification (or categorization) is a machine learning problem which gets more and more important day by day. Since beginning for the internet in the 90's, the number of internet user s and the number of web paging serving to have increased at such a rapid pace and continue to grow.

II. Introduction of Web Mining

Web mining is an application of data mining techniques to find information patterns from the web data. Web mining helps to improve the power of web search engine by identifying the web pages and classifying the web documents. Web mining is very useful to e-commerce websites and e-services. The contents of data mined from the Web may be a collection of facts that Web pages are meant to contain, and these may consist of text, structured data such as lists and tables, and even images, video and audio. Web mining can be broadly divided into three different types of techniques of mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it..

III. Web Page Classification

3.1 Introduction of Web Page Classification

Web pages classification (or categorization) is a machine learning problem which gets more and more important day by day. Since beginning for the internet in the 90's, the number of internet user s and the number of web paging serving to have increased at such a rapid pace and continue to grow.

3.2 Types of Classification

Based on the number of classes available, a classification problem can be divided into a binary classification in which instances should belong to one of two classes, and into a multiclass classification where more than one class is defined. When only one label is assigned to an instance, the classification problem is defined as single-label classification. But if more than one class is assigned to an instance, the classification is then referred to as multilabel

DOI: 10.48175/568

one. We can also divide web page classification into flat and hierarchical classification where categories are parallel in the former and organized in a hierarchical tree structure in the latter, in which each category may have several subcategories.

3.3 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.

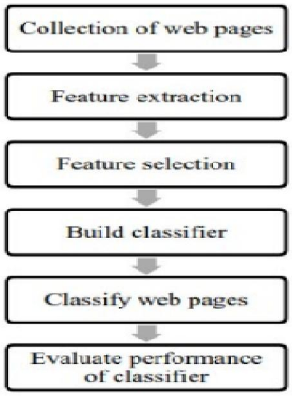

Figure 1: Webpage Classification Process

A. 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.

B. 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.

C. 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. 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.

D. 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.

E. 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.

F. Evaluate Performance Of Classifier

The metrics used to evaluate the performance of classifier are precision, recall, F-measure and accuracy.

3.4 Background and Related Work

Before reviewing Web classification research, we first introduce the problem, motivate it with applications, and consider related surveys in Web classification.

A. Problem Definition

The increasing number of web pages has caused some problems which it is difficult for the user to get relevant result, malicious unsafe web pages also there, when searching information in the search engine. Thus to solve this problems, the web page classification is needed.

B. Related Survey

A deep learning-based system has been developed for the classification of web pages in this study. The data obtained from Roksit was used in the developed system. The meta tag information contained in the web page is used to classify a web page. The meta tags used are title, description and keywords. This information has been collected by the crawler module developed in this study. RNN based deep learning architecture was used during the tests. The effect of using transfer learning on the system has also been examined. The tests were performed on a leased GPU. According to the results obtained, success rate of web page classification system is approximately 85%. It is not observed that transfer learning has significant contribution to the success rates.

Table 1: Comparisons of Literature Survey

Paper	Method	Advantages	Disadvantages
A survey on technique for solving web page classification problem	Convolution Neural Network (CNN)	High classification accuracy Reduce the computation time	Does not deal with redundant features Increased runtime
Web page Classification using RNN	Recurrent Neural Network (RNN)	Work more efficiently Time complexity O(n)	The RNN architecture tend to have bias
Web Mining for Information Retrieval	Support Vector Machine (SVM) Navies Bayse	Useful information can be easily predicted	Complex data which can uses different algorithms
Web Page Classification Method Based on Semantics and Structure	Support Vector Machine (SVM) Convolution neural networks (CNN)	Useful structural and semantic information	Less size of corpus Complicated
Machine Learning for Web Page Classification: A Survey	K-nearest Neighbor Support Vector Machine (SVM) Navies Bayse	Decreases the noisy elements on the web pages Neural networks work better than other methods	Interpretation of results
Web page classification: a survey of perspectives, gaps, and future directions	Text-Based Classification Image-Based Classification Combination of Text and Image-Based Classification	Multiview representations	Low accuracy Extracting complicated hand-crafted features
Prediction of user's type and navigation pattern using clustering and classification algorithms	Path prediction Page gathering Fuzzy clustering Ant-based clustering Graph partitioning	fuzzy clustering gives higher prediction accuracy	Clusters are more overlapped

DOI: 10.48175/568

2.5 Result of Work

In this study, the information obtained from the web pages was used to classify the web pages. There are also studies where information about neighboring web pages is used to classify a web page. Meta tags contain information about the purpose and content of a web page. In this study, meta tags were used to classify web pages. The meta tags used are title, description and keywords. RNN architecture was used during the tests. Achieved success rate is approximately 85%. (TL: Transfer Learning was used. - Non-TL: Transfer Learning was not used.)

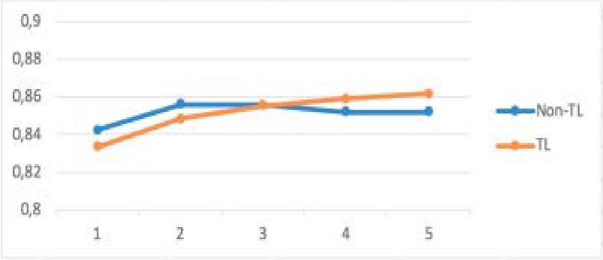

Figure 2: Validation Accuracies for the Comparison of TL and Non-TL

III. Proposed Methodology

In this paper we proposed Web page classification technique. Still although content based topic classifiers gave enhanced consequences than URL-based ones, concern classification from URL is preferable when the content is not available, or when classification has the major significance. We can findings for URL-based Web page topic classification as follow.

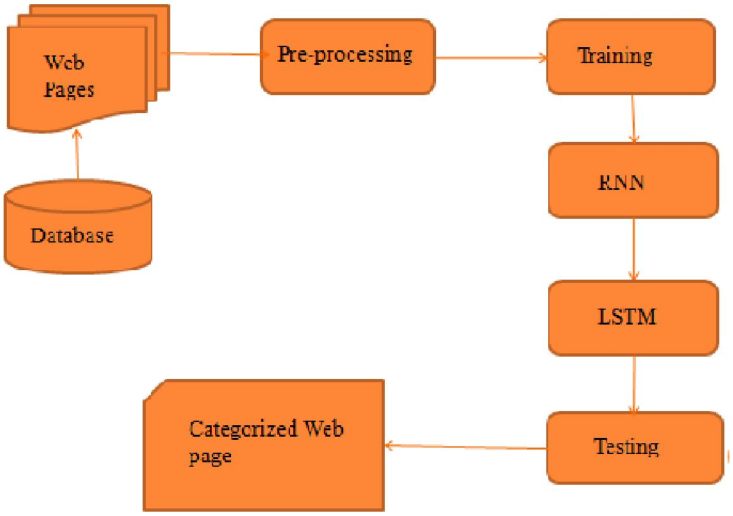

Figure 3: Proposed System Methodology

3.1 Pre-processing

Pre-processing is a process to convert raw data into meaningful data using different techniques. In this proposed system, get the input URL from the database which created and performed pre-processing phase. In pre-processing phase we used data transformation technique to select the attribute.

3.2 Training Phase

This module is for extracting relevant information on training dataset of URLs and storing the information in a binary file (pickle file).

DOI: 10.48175/568

3.3 Recurrent Neural Network

Recurrent neural networks (RNN) are a class of neural networks that are helpful in modelling sequence data. Derived from feed forward networks, RNNs exhibit similar behaviour to how human brains function. Simply put: It produce predictive results in sequential data that other algorithms can't. In RNN, the output from previous step is fed as input to the current step. In the RNN architecture, while a word at the step of t is processed, the information of the words before the step of t is also taken as input. The basic RNN architecture consists of cells that are repeated one after the other. A cell is taken previous cell information and given word as inputs. The recursive representation is plotted over a single cell in some references. Some other references represent the architecture as the sequential cells.

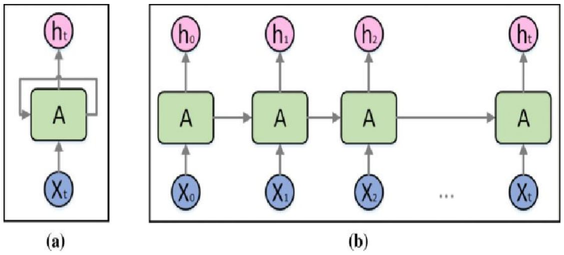

Figure 4: (a) RNN architecture (b) An unfolding structure of RNN

3.4 Long Short Term Memory

Long Short Term Memory networks are a variant of recurrent neural networks (RNN). LSTM is used to learn some context-dependent information from sequential data, for example, some information related to text inputs or video data. It is an enhancement of standard RNNs in terms that LSTM can easily learn long-term dependencies, which is not possible in standard RNN. This makes LSTMs much more human like, since humans also use contextual information to predict future data. LSTMs can remember past information over a long period of time and this property makes it useful for many applications.

3.5 Categorized Web Pages

In this way, the URL based classification is performed using RNN with LSTM. The final output layer is the dense layer with the softmax activation function, which is used to decide whether the given URL is safe or malicious link affected like phishing, spamming.

IV. Conclusion

This technique presented an proficient technique for web page classification. This technique added effective is the training set is set in such a technique that it produce added sets. Though the experimental consequences are relatively encouraging, it would improve if the work with superior data sets with added classes. The existing techniques require added or less data for training as well as less computational time of these techniques.

Acknowledgment

I am thankful to all those who directly or indirectly help me in accomplishing my task. Last, but not the least I am grateful thank my institute, Atmiya University for giving me this opportunity to work in the great environment.

References

[1] Ebubekir buber, Banu diri, "Web page Classification using RNN ". International congress of information communication technology-ICICT, 2019.

[2] Siti Hawa Apandi, Jamaludin Sallim and Rozlina Mohamed "A survey on technique for solving web page classification problem", IOP publishing, 2020.

[3] Mahdi Hashemi "Web page classification: a survey of perspectives, gaps, and future directions". Springer Science+Business Media, LLC, part of Springer Nature, 2020

[4] Huaxin Li , Zhaoxin Zhang, Yongdong Xu "Web Page Classification Method Based on Semantics and Structure", 2nd International Conference on Artificial Intelligence and Big Data-IEEE ,2019.

[5] Safae lassri, EL HABIB BENLAHMAR, Abderrahim TRAGHA "Machine Learning for Web Page Classification: A Survey", International Journal of Information Science & Technology-IJIST ,2019, Vol. 3 - No. 5.

[6] R. Rajalakshmia, Hans Tiwarib, Jay Patelb, Ankit Kumarb, Karthik.R "Design of Kids-specific URL Classifier using Recurrent Convolutional Neural Networ" International Conference on Computational Intelligence and Data Science (ICCIDS)2019.

[7] Shadab Irfan , Subhajit Ghosh , "Web Mining for Information Retrieval" , International Journal of Engineering Science and Computing-IJESC, 2018, Vol-8, No-4.

[8] D. Anandhi, M. S. Irfan Ahmed, "Prediction of user's type and navigation pattern using clustering and classification algorithms" Springer Science+Business Media, LLC, part of Springer Nature,2017.

[9] Neeraj Mehta, Avinash Rathore "New technique for Web page Information Categorization using Unsupervised Clustering" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (2) , 2016.

[10] S.Vidya, K.Banumathy "Web Mining- Concepts and Application" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015.

DOI: 10.48175/568