Accepted Manuscript

Machine learning based heterogeneous web advertisements detection using a diverse feature set

Ab Shaqoor Nengroo, K.S. Kuppusamy

PII: S0167-739X(17)32877-7
DOI: https://doi.org/10.1016/j.future.2018.06.028
Reference: FUTURE 4290

To appear in: *Future Generation Computer Systems*

Received date : 15 December 2017
Revised date : 31 March 2018
Accepted date : 18 June 2018

Please cite this article as: A. Shaqoor Nengroo, K.S. Kuppusamy, Machine learning based heterogeneous web advertisements detection using a diverse feature set, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.06.028

Machine Learning based Heterogeneous Web Advertisements Detection Using a Diverse Feature Set

Ab Shaqoor Nengroo, K S Kuppusamy*

[a]*Department of Computer Science, School Of Engineering and Technology, Pondicherry University, Pondicherry 605014, India*

Abstract

Advertisement identification and filtering in web pages gain significance due to various factors such as accessibility, security, privacy, and obtrusiveness. Current practices in this direction involve maintaining URL-based regular expressions called filter lists. Each URL obtained on a web page is matched against this filter list. While effectual, this procedure lacks scalability as it demands regular continuance of the filter list. To counter these limitations, we devise a machine learning based advertisement detection system using a diverse feature set which can distinguish *advertisement blocks* from *non-advertisement blocks*. The method can act as a base to provide various accessibility-related features like smooth browsing and text summarization for persons with visual impairments, cognitive impairments, and photosensitive epilepsy. The results from a classifier trained on the proposed feature set achieve 98.6% accuracy in identifying advertisements.

Keywords: advertisements, web accessibility, content extraction random forest, machine learning

*Corresponding author

Email addresses: shakoor.ab.phd.cse@gmail.com (Ab Shaqoor Nengroo), kskuppu@gmail.com (K S Kuppusamy)