World Academy of Science, Engineering and Technology
International Journal of Information and Communication Engineering
Vol:12, No:03, 2018

An Integrated Lightweight Naïve Bayes Based Webpage Classification Service for Smartphone Browsers

Authors : Mayank Gupta, Siba Prasad Samal, Vasu Kakkirala

Abstract : The internet world and its priorities have changed considerably in the last decade. Browsing on smart phones has increased manifold and is set to explode much more. Users spent considerable time browsing different websites, that gives a great deal of insight into user's preferences. Instead of plain information classifying different aspects of browsing like Bookmarks, History, and Download Manager into useful categories would improve and enhance the user's experience. Most of the classification solutions are server side that involves maintaining server and other heavy resources. It has security constraints and maybe misses on contextual data during classification. On device, classification solves many such problems, but the challenge is to achieve accuracy on classification with resource constraints. This on device classification can be much more useful in personalization, reducing dependency on cloud connectivity and better privacy/security. This approach provides more relevant results as compared to current standalone solutions because it uses content rendered by browser which is customized by the content provider based on user's profile. This paper proposes a Naive Bayes based lightweight classification engine targeted for a resource constraint devices. Our solution integrates with Web Browser that in turn triggers classification algorithm. Whenever a user browses a webpage, this solution extracts DOM Tree data from the browser's rendering engine. This DOM data is a dynamic, contextual and secure data that can't be replicated. This proposal extracts different features of the webpage that runs on an algorithm to classify into multiple categories. Naive Bayes based engine is chosen in this solution for its inherent advantages in using limited resources compared to other classification algorithms like Support Vector Machine, Neural Networks, etc. Naive Bayes classification requires small memory footprint and less computation suitable for smartphone environment. This solution has a feature to partition the model into multiple chunks that in turn will facilitate less usage of memory instead of loading a complete model. Classification of the webpages done through integrated engine is faster, more relevant and energy efficient than other standalone on device solution. This classification engine has been tested on Samsung Z3 Tizen hardware. The Engine is integrated into Tizen Browser that uses Chromium Rendering Engine. For this solution, extensive dataset is sourced from dmoztools.net and cleaned. This cleaned dataset has 227.5K webpages which are divided into 8 generic categories ('education', 'games', 'health', 'entertainment', 'news', 'shopping', 'sports', 'travel'). Our browser integrated solution has resulted in 15% less memory usage (due to partition method) and 24% less power consumption in comparison with standalone solution. This solution considered 70% of the dataset for training the data model and the rest 30% dataset for testing. An average accuracy of ~96.3% is achieved across the above mentioned 8 categories. This engine can be further extended for suggesting Dynamic tags and using the classification for differential uses cases to enhance browsing experience.

Keywords : chromium, lightweight engine, mobile computing, Naive Bayes, Tizen, web browser, webpage classification
Conference Title : ICWS 2018 : International Conference on Web Services
Conference Location : Madrid, Spain
Conference Dates : March 26-27, 2018