A Machine Learning Approach to Improve the Performance of Web Proxy Cache Replacement

S. Nimishan and S. Shirparen

Department of Computer Science, Faculty of Science, University of Jaffna

INTRODUCTION

Web caching is a well-known strategy for improving the performance of Web proxy servers. Proxy servers play a key role between users and Websites in reducing the response time and saving network bandwidth. As shown in Fig 1, the proxy cache is found in the proxy server, which is located between the client machines and origin server.

Due to cache space limitations, a Web proxy cache replacement policy is required to manage and manipulate the Web proxy cache contents efficiently and effectively.

METHODOLOGY

Proxy cache miss occurs in a web client communication with the proxy server, if the requested object is not in the proxy cache or not fresh. The proxy cache manager needs an approach to know which objects to keep or which objects to remove that is the unwanted Web objects in order to release enough space for the new coming object. This approach is to use machine learning classifiers to evict unwanted web objects from the early stage before caching.

In the classification of log files, unwanted attributes in the log are removed and a new attribute as target attribute is added. That will have a value of ‘1’ if the domain name re-requested again within the 30 minutes Sliding Window Length (SWL), otherwise it will be assigned a value of ‘0’. Then the trace set is ready to be applied under the Machine Learner approaches using the WSO2 ML and therefore divide the dataset to take 70% as data training and 30% as test data.

Data pre-processing requires two steps: log file preparation and training dataset preparation. In the log file preparation, irrelevant or invalid requests are removed from log files such as un-cacheable web requests.

PERFORMANCE EVALUATION

We can improve the accuracies by removing the unwanted attributes such as size of the web object requested from the dataset. Because there are few fields that are not used for classification.

The hit ratio (HR) is the most widely used metrics for evaluating the performance of Web caching. HR is defined as follows:

\[
HR = \frac{\text{Number of Hits}}{\text{Total Requests}}
\]

Regarding Decision Tree training, the settings and default values of parameters as determined in WSO2 ML were changed according to our requirement.

This study considers that each Web object belongs to the positive class if the object is re-requested again. Otherwise, the Web object belongs to the negative class.

DATA COLLECTION AND PRE-PROCESSING

The data have been downloaded from Billion Triples Challenge 2012 Dataset as raw log entry files which contains 4,751,262 entries. In the dataset and there are no HITs or TCP HITs in the http rule code column. Each line in the proxy logs file represents access proxy log entry, which contains the following ten fields: Timestamp, Etagged Time, Client IP address, Log tag and HTTP code, Size, Request Method, URL, User identification, Hierarchy Data and Hostname, and Content type.

CONCLUSION

In the pre-processing step unsuccessful HTTP requests are eliminated and have included the successful HTTP requests that had a states code of 200. Also out of the totally collected 4,751,262 entries only 39.73% were the cacheable requests.

REFERENCES


SLWSCS - 2017