

# ONLINE ADVERTISING: FORECASTING AND SYNTHESISING WEB ACTIVITY BASED ON HISTORICAL DATA

*Pedro Manuel Santos Borges*

Dissertation conducted under the supervision of Prof. João Moreira, co-supervision of Prof. Hugo Ferreira  
and company supervision of Eng. João Azevedo  
at *ShiftForward, Lda*

---

## 1. Context and Framing

The online marketing is a growing multibillion-dollar industry [1] which is expected to continue its fast growth[2].

This industry is always trying to become more efficient by getting more profit from assets it already owns. Web users are the major assets of this industry, which makes money by exploiting the user behavior and characteristics, in order to target them with the perfect campaign. Each campaign has its own target parameters, which limit the target user universe. Online marketing industry core business is centered in web users and this industry has recorded almost every footprint each user makes on the web. Future footprints of the web users allows the measurement of the behavior of an upcoming campaign and, with this data, it is possible to make the inventory more profitable. Therefore, using future user data allows the adtech industry to be able to fine tune its campaigns. Campaigns are composed by a series of ads that share the same main idea they want to transmit. The campaigns have a targeting typically defined as a set of parameter definitions and rules. To be able to run the campaigns in a simulator, their targeting can be defined as queries over the ad requests' data. The utilization of a simulation allows to get the results fast, test concurrent campaigns and test multiple scenarios. The utilization of a simulation is the main reason behind why is so important to be able to generate future ad requests' data.

The most common platforms that will benefit from this data are Custom-Built Ad Servers and Exchanges, Sell-Side Platforms (SSPs) and Demand-Side Platforms (DSPs).

## 2. Project

Our goal is to forecast the availability of the users in the future so we can simulate the value of future campaigns over them.

Since we do not know which characteristics the future campaigns will have, every detail available on the impressions needs to be forecasted, in order to obtain the correct values when the queries are executed over the generated data.

With this said, the main goal is to be able to generate a dataset able to be used on a campaign simulator,

so we must be able to predict the values with the maximum detail possible.

This simulator needs to have available every detail possible about every impression in order to identify which impressions are compatible with each campaign. The result would be expressed by number of impressions per campaign and users target by every campaign, over time.

The approach should also only need an impression's date and user id, with all other variables being optional. This constraint is imposed by the multiple sources of the dataset used, since each one of these sources could store different details and different types of parameters about each impression, so we cannot rely on the availability of such parameters.

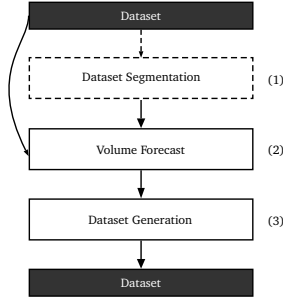
## 3. Motivation and Goals

In the last few years, the online marketing has been getting more complex. In such a way that today campaigns have a very well defined target, with sets of rules and limitations. This poses a big problem to simpler prediction models that normally don't predict all the parameters of the ad request, this way limiting the parameters where queries can be done.

Nowadays, some online ads can only be imprinted if a set of very specific requirements has been fulfilled, for example, the users had to visit an e-commerce site in the last 24 hours. This brings causality into the equation, creating a new paradigm that makes the more traditional methods of prediction ineffective. To solve this problem and to be able to get fast responses to complex queries of concurrent campaigns, simulate the algorithms executed by ad servers of the client and ultimately parallelize the computation of the results for the queries, the complete future data has to be predicted. This generated data can be used in simulations and the online campaigns can run on top of the future population.

The objective of this thesis is to develop a library capable of generating future ad request logs using past data from the same network. This library will have as one of its main goals the prediction of all the parameters that characterize an ad request with the purpose of being able to query over any parameter, in other words, the generated dataset (ad requests log) must have the same attributes as the original.

## 4. Approach



**Fig. 1 – High level overview of the approach**

As the figure 1 makes clear, the goal of the proposed approach is to use a dataset containing logs of the web activity from an online advertising related network and use this information to generate a possible future web activity logs on the same network. This should be done in order to preserve tendencies and into data coherency.

This approach can be divided into three main phases:

1. *Segmentation*, which its main purpose is divide the dataset in smaller and more predictable datasets, in order to improve the results obtained after the second phase, mostly when there are large quantities of data available.
2. The second phase is where the *forecast of the volumes* that characterize the traffic on the network are done, using time series prediction methods (experiments were made using *ARIMA*).
3. The third and last phase of the process, the more complex one, is where the volumes generated from the phase two combined with the data provided by the original dataset are used to *generate a dataset* that represent a possible future of the web activity on the target network.

## 5. Results

	$\sigma$ (Real Data)	RMSE	MASE
without segmentation	2224.74	1237.99	0.1958
baseline (100 segments)	2224.74	1352.42	0.2174
segmentation per parameter (browser)	2224.74	1226.60	0.1884
segmentation data-stream clustering (threshold 20)	2224.74	1786.43	0.2927

**Tab. 1 – Impression volume forecasting on a real dataset**

	$\sigma$ (Real Data)	RMSE	MASE
without segmentation	6.70	9.24	1.2511
baseline (100 segments)	6.70	6.88	0.7896
segmentation per parameter (browser)	6.70	10.75	1.3883
segmentation data-stream clustering (threshold 20)	6.70	15.93	2.5960

**Tab. 2 – Impression volume forecasting on real dataset, queried by "Portugal"**

In about every test case, the results that used the segmentation phase got better results, when compared to a non-segmented approach. (some examples on Table 1 and 2) These results were only compared to a small number of queries, so the results can be completely different when used in its goal environment, the online advertising campaigns impact on that particular network.

## 6. Objectives Fulfilment

We proposed the use of the three phase approach. The first phase allows better predictions for certain characteristics, the second phase allows the prediction of volumes in the future, and lastly the third phase fills the predicted future data with additional information from the past while following the tends of the predicted volumes. Our results show that the usage of dataset segmentation returns good to moderate improvements while predicting the volumes.

## 7. Future Work

In order to achieve better results, some other approaches for each phase could be explored. Other algorithms of segmentation phase for example the usage recursive parameter segmentation. For example, segment the whole dataset by browser and then by country. It is my believe that this approach would get good results if a huge volume of data were available. On the volume prediction other algorithms and techniques like *MLP*, multilayer perceptron, could be used. On the dataset generation phase, the knowledge for the constitution of some attributes and additional volume paramters (example, percentage of new cities) could help improve some of the results.

## References

- [1] I.A.B. PricewaterhouseCoopers. Iab internet advertising revenue report, 2012 full year results, April 2013.
- [2] I.A.B. PricewaterhouseCoopers. Q3 2013 internet advertising revenues climb to landmark high of nearly \$10.7 billion, marking 15% year-over-year growth, December 2013.