URL Classification Project Report

Author: Sujai Hiremath

Collaborators: Shunto Kobayashi, Miguel Angel Alcobendas Lisbona, Matthew S. Shum

Date: Oct 27, 2021

---

1. Motivations

Amongst our collaborators, there are at least two different projects underway that are relevant to this project. Shunto, Miguel and Matt have been working on a paper on the impact of privacy measures on online advertising auctions [1]. They and I have been working on a separate project about the value added by bidders known as rebroadcasters in online ad auctions. Both projects involve websites bidding on the opportunity to advertise themselves through intermediaries. However, information about the types of companies engaging in bidding is sparse, as such company category data is self reported. Such information would be useful towards understanding the bidding behavior of different industries online, and would aid both aforementioned projects above in their pursuits. Therefore, this project attempts to create a machine learning model to assign industrial category labels to companies using their website URL and potentially other methods such as web scraping.

2. Goals

The overarching goal of this project is to build a tool that allows us to efficiently and accurately classify large data sets of website URLs into their respective industrial categories. We break this down into the following sub goals:

1. Obtain and clean a data set of website URLs that have been accurately labelled with categories.
2. Identify a suitable method for classification given our constraints through literature review.
3. Implement the method of classification.
4. Iterate to improve the model.

3. Work Accomplished

I attach all relevant code at the end of the report.

Shunto reached out to Miguel to obtain a Yahoo! data set of website URLs categorized by industry. He took the raw data and formatted it into a usable form, cleaning it substantially. From there, I helped clean the data by removing multi-classified websites and non-English websites.

Shunto and I both conducted a brief literature review, reading through relevant papers on URL classification. We ended up basing our method on [2], a paper I found. This method chiefly involves (amongst other things) processing website URLs into 3 letter ngrams and applying a machine learning algorithm for classification.

I used the tidymodels package in R to create a machine learning model that processes website URLs into 3 letter ngrams, and uses a XGBoost classifier algorithm to assign category labels. The model currently has a 95.25% accuracy.

4. Future Work

I have been, with the advice of Shunto, working on changes to the algorithm to improve its accuracy. We focus on 3 main changes:

1. Generating more raw data (such as website text) to feed into the algorithm.
2. Processing the data in different ways to make analysis more efficient.
3. Tweaking the settings/parameters of the algorithm to yield better accuracy while avoiding overfitting.

In particular, I was only able to use the R package Rvest to implement web scraping in small data sets. I aim to make this work soon.

References

[1] Alcobendas, Kobayashi, Shum, *The Impact of Privacy Measures on online Advertising Markets.*

[2] Berardi, Esuli, Fagni, Sebastiani, *Classifying websites by industry sector: a study in feature design.*