A COMPARISON OF URL FINDERS FOR ONLINE-BASED ENTERPRISE CHARACTERISTICS

V. Nekrašaitė-Liegė[1]

[1] Vilnius Gediminas technical University, Statistics Lithuania, Lithuania
e-mail: vilma.nekrasaite-liege@vilniustech.lt

Abstract

One of the fields where the integration of big data in the regular production of official statistics might be possible is Online-based Enterprise Characteristics. Currently two URL finder programs are suggested by ESSnet, thus overview and comparison of these two URL finders will be presented here.

Keywords: Online-based Enterprise Characteristics, URL finder.

1 The web as a statistics data source

Web scraping is easy, however if you want to use it as a statistics data source, you want it to be automated, methodologically sound, transparent, robust, consistent and efficient. For this reason the ESSnet Web Intelligence Network project was created and Statistics Lithuania is a member of it. The project goal is to contribute to establishing the Web Intelligence Network (WIN) across the ESS and to make use of the Web Intelligence Hub (WIH) services for the production of statistics with web data. This project started in 2021 and is going to last 4 years. Currently, there are two main fields (Online Job Advertisements (OJA) and Online-based Enterprise Characteristics (OBEC)) where the initial steps are made. This article will focus on the work done in the OBEC field.

The use of OBEC data would support the official statistics with more recent data, it would improve the Statistical Business Register (SBR) and could be used as additional information in Information and Communication Technology (ICT) surveys.

The first OBEC goal is to create a database containing URLs for each enterprise in the target population, where there can be one, many or zero URLs for a given enterprise. For some enterprises the URLs might already be available in a SBR or obtained from other sources. It can also be built up from scratch by searching for enterprises via search engines like Bing or Google. Of course, web scraping can be used as a verification tool. Thus, the search results can help to answer two main questions:

- Does an enterprise have a website?
- Which URL is most likely to belong to that enterprise?

Commonly, a web search will return several results leading to different base URLs for one enterprise. The different machine learning methods and algorithms like logistic regression or random trees can be used to identify a valid URL.

The previous projects (ESSnet Big Data I and ESSnet Big Data II) also investigated this field and two URL finder softwares were created:

- **UrlSearher**: `https://github.com/SummaIstat/UrlSearcher`

- **URLsFinder**: `https://github.com/EnterpriseCharacteristicsESSnetBigData/StarterKit/tree/master/URLsFinder`

A more detailed analysis of these programs is provided in the next section.

2 Comparison of URL finders

UrlSearher was created by Donato Summa and his team (Italian National Institute of Statistics) (Barcaroli G., Scannapieco M., Summa D. (2016)). UrlSearcher is a Java application where a strategy for solving the URL retrieval problem is adopted. It consists of 5 steps:

- *Step 1: Building the input training dataset.* Combining different sources the list of enterprises with several indicators (enterprise name, city, telephone number and ect.) is created. This step must be done outside the URL finder, because each country can use different sources and different indicators and there is no possibility to automate this process.
- *Step 2: URLs Searching.* In this step for each unit in the input training dataset the first 10 URLs were stored from the search engine, where the search was done using the enterprise name.
- *Step 3: URLs Crawling.* For each row of the seed file, if the URL is not in the list of the domains to filter out, the program tries to acquire the HTML content of the page. From each acquired HTML page the program extracts just the textual content of the HTML fields with useful information (for example, contact information) and write a line in a TSV file.
- *Step 4: URLs Scoring.* A score vector is computed and a score is assigned for each line in the TSV file. The elements/characteristics that were considered in a score vector by default (it is possible to adapt it to each country) are these:
  - Simple URL (is the URL in the form www.name.lt or not?);
  - VAT (is it present in the page or not?);
  - city (is it present in the page or not?);
  - province code (is it present in the page or not?);
  - link position (from 0 to 9);
  - telephone number (is it present in the page or not?);
  - zip code (is it present in the page or not?).

  A score is calculated as a sum of assigned points for each element/characteristic.
- *Step 5: Using a Machine Learning approach to associate URLs to enterprises.* The easiest way to assign the valid URL for each unit is to select that with the maximum score, but knowing that not all units have a URL, the more precise algorithm must be used. That is why in this step three methods (neural networks, random forest and logistic model) are used to determine if the URL with the highest score is valid or not.

The other program **URLsFinder** was created by Kostadin Georgiev (Bulgarian National Statistical Institute) and is a part of a Starter Kit package. The URLsFinder is written in Python and it contains two main modules:

- *URLsFinderWS* - defines methods for scraping information for the enterprises' URLs from the internet with the help of search engine Duck Duck Go.

- *URLsFinderMLLR* - defines methods for determining the enterprises' URLs or characteristics from the scraped information from the internet by using logistic regression machine learning.

As the UrlSearher, the URLsFinder has a similar course of action, still there are some differences, which are presented in table 1.

Table 1: A comparison of URL finders

	UrlSearher	StarterKit
Language	Java	Python
Search engine	Bing	Duck Duck Go
Characteristics included in a score vector (by default)	Simple URL VAT city province code link position telephone number zip code	Simple URL ID city address link position telephone number name equal domain
Machine Learning methods	neural network random forest logistic	logistic

More detailed comparison and adaptation to Statistics Lithuania needs will be presented during the presentation.

3 Some observations

ESSnet WIN project is still at the early stage, thus the main results will be obtain in the future, however, some observations regarding OBEC field can be already made:

- Even if we agree that the web scraping is a powerful tool to obtain the information, still at this moment it won't change the traditional survey sampling, but it can provide useful up-to-date additional information, which could be integrated in the survey sampling procedures.
- It is necessary to define country specific steps and stages for collecting the data, thus the programs must be easily updated.
- To validate that suggested URL is correct the machine learning methods are used, where there is a need to have a train set. Unfortunately not always there is a possibility to construct an appropriate train set.

References

Barcaroli G., Scannapieco M., Summa D. (2016) On the use of internet as a data sourse for official statistics: a strategy for identifying enterprises on the web. *Rivista Italiana di Economia Demografia e Statistica*, **LXX**, 25-41.

ESSnet Big Data I. WP2 led by Monica Scannapieco/ISTAT (OBEC) https://ec.europa.eu/eurostat/cros/content/wp2-webscraping-enterprise-characteristics_en

ESSnet Big Data II. WPC led by Galia Stateva/BNSI (OBEC) https://ec.europa.eu/eurostat/cros/content/WPC_Enterprise_characteristics_en