Methodology overview

Enterprise Characteristics Experimental Statistics Netherlands

Date: 31-10-2019

Disclaimer: *The ESSnet BD WPC explores new data sources for determining enterprise characteristics using as much as possible the same methodological concepts in multiple countries. It should be noted that these statistics have not reached maturity in terms of harmonisation, coverage or methodology. At this point they are to be treated as experimental indicators and they do not necessarily align with the official statistics published on this subject.*

Aims

Statistics Netherlands performed multiple research projects using multiple data new sources to determine characteristics of enterprises from web data. In this ESSnet work package we combine some of these approaches to calculate experimental statistics on the use-cases defined in such way that it aligns as much as possible with the work performed by the International colleagues in the joint project. The population that was agreed upon was the population of enterprises as defined by the ICT use survey (>10 persons, limited NACE). In this wave of experimental statistics it was agreed to focus on the following indicators:

- Use-case 1 URLs Inventory
    - Rate of enterprises having websites
        - Enterprises by NACE categories
        - Enterprises by NUTS Level 3
- Use-case 2 Variables in the ICT usage in enterprise survey
    - E-commerce – Rate of enterprises engaged in web sales on their website
    - Social Media Presence – rate of enterprises that are present on social media
    - Rate of enterprises having specific features of the website
        - Advertisement of open job positions or online job application

Data Sources

Statistics Netherlands used the following data sources to produce these experimental statistics about the enterprises:

- Statistical Business Register (SBR)
    - About 1.6 M enterprises of which 1/3 have a known URL
    - Some of the variables used: Enterprise ID, Name, chamber of commerce ID, known website (if available), address, number of employees, NACE
- Google Custom Search API
    - Six search requests per enterprise
    - First 10 search results of each query

- Enterprises web sites
    - Scraping texts from web pages where needed.
- Dataset from DataProvider
    - A dataset from the Dutch company dataprovider.com. This company indexes the web and structures data found to gain insights about companies. The dataset used is from the Dutch .nl domain.

Methodology

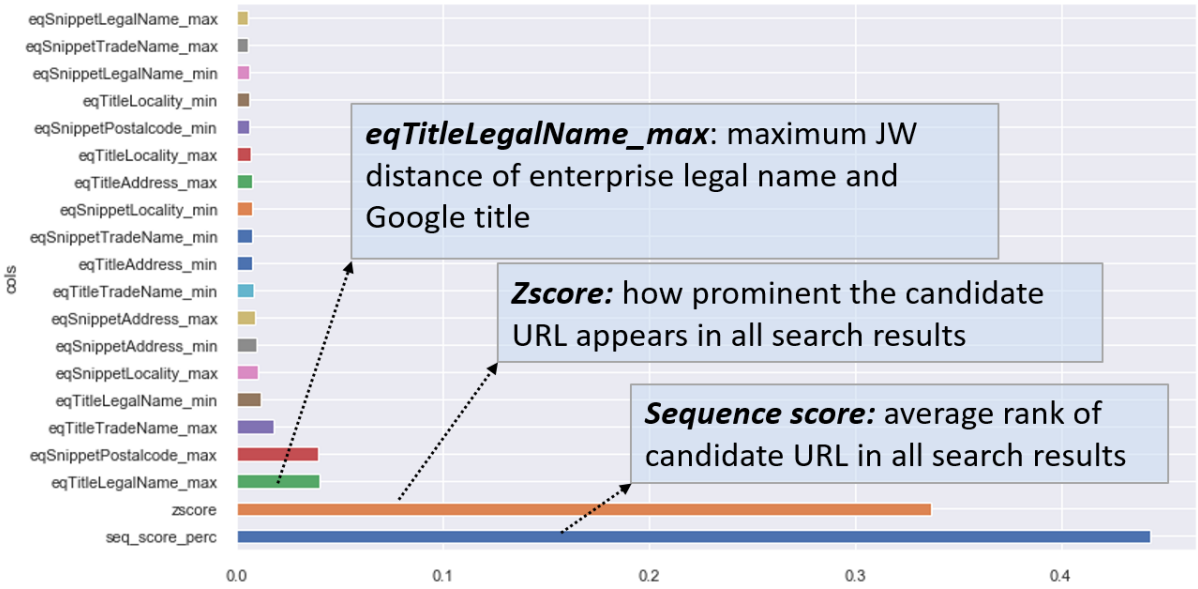
Use-case 1 URLs Inventory:

Training phase:

1. The model used in the previous ESSnet project to determine whether a found URL is correct has been improved further in this project. The training set was enlarged from different data sources. It now consists of 3300 enterprises *with* URL from the SBR, > 10 persons, any NACE code.
2. For all enterprises in this set, we executed 6 search queries on Google, using the Google Custom Search API. Five queries use a combination of enterprise name and address info. One query is designed to find the contact page of an enterprise. The queries are:
    - Name
    - Name + 'contact'
    - Name + Street + 'contact'
    - Name + PostalCode + 'contact'
    - Full address
    - Name + 'inanchor:contact'

    For each query the first 10 results were used.

3. We split this set into a training set and a validation set and created a model (SVM) to predict if a URL from the search results belongs to the enterprise being searched for. About 30 features are defined on all Google search results. For example a feature checks whether the name or postal code is contained in the URL, title or snippet. Also the rank in the search results list and the type of query is used as a feature. The figure below shows the feature importance of each feature in the model and gives the explanation of the three most important features. For a more extensive explanation of the methodology behind this model we refer to [1] and [2] (to appear).



Prediction phase:

In the datasets that are available to Statistics Netherlands there is already a lot of data available, retrieved from the web, that can be used for calculating experimental statistics. For many enterprises the URL is known from additional data sources. Thus, the URL retrieval model has to be applied only to the enterprises for which we do not have a URL from any of these data sources. Therefore at first, a *baseline* of enterprise characteristics was created from the SBR, linking it to the data from DataProvider and the results of some previous URL finding exercises in which the model described above was used. From this set a subset was created consisting of the enterprises that belong to the population of the survey on ICT use (> 10 persons, limited NACE) conforming to the agreed population definition in this work package. This so called ICT baseline dataset was used to calculate the experimental statistics. Enterprises for which no website has been found in any of the data sources are supposed to not have a website. If the dataset contains at least one website for the enterprise it is supposed to have a website.

A complicating issue in the approach above is the transition from legal units to enterprise units. The linking of multiple datasets has been performed on the level of legal units, while the experimental statistics are to be published on the level of enterprises. It could be that a certain variable has a different value for different legal units belonging to the same enterprise. In the transition from the legal unit level to the enterprise level, we took a practical approach. An enterprise is supposed to have a website if this holds for at least one of its legal units. Vice versa for the other experimental indicators. For a further explanation of this issue we refer to [2].

Use-case 2 Variables in the ICT usage in enterprise survey

E-commerce – Rate of enterprises engaged in web sales on their website

The ICT baseline dataset contains multiple variables on ecommerce from the DataProvider dataset which were derived from scraping websites for which a URL is known. After some experimentation with these variables we found that the variable 'eCommerce certainty' could well be used to calculate experimental indicators on web sales. Since this variable expresses a certainty on E-commerce a threshold had to be chosen. This was somewhat arbitrary. Playing with different threshold values showed higher or lower values for the experimental indicators, but it also showed that similar distributions over size classes, NUTS3 and NACE. A value of .5 was chosen for the calculation of the experimental rates of enterprises engaged in web sales on their website, which are expressed in the results, the total, by size class, by NUTS3 and by NACE.

Social Media Presence – Rate of enterprises that are present on social media

The ICT baseline dataset contains multiple variables on social media presence on websites for enterprises for which a URL is known. The social media possibly detected Facebook, Twitter, LinkedIn, Instagram, YouTube and Pinterest. For these experimental indicators we defined an enterprise to be present on social media if it scores positive on either one of these social media channels. Using this definition we calculated the experimental rate of enterprises present on social media via their website, the total, by size class, by NUTS3 and by NACE in the tables provided.

Rate of enterprises having specific features of the website

Unfortunately the ICT baseline dataset that we created in this exercise did not contain any usable variables on the advertisements of open job positions or online job application on websites of enterprises. Either additional scraping would be necessary or alternatively we could search for another dataset containing such variables to link to the ICT baseline dataset. Knowing that multiple projects exist studying job vacancies (i.e. the European CEDEFOP as well as national initiatives), we choose at this moment not to do additional scraping in this work package but instead to look for possibilities to link other datasets resulting from scraping to the baseline for this particular goal. This will be studied in more detail in the sequel of the work in WPC.

Results

The experimental statistics calculated using the methods described in this document are available on the WPC wiki page of the work package, under Netherlands:

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WPC_Experimental_statistics

The experimental indicators can be found in table 1 (total), table 2 (by size class), table 4 (by NACE) and table 4 (by NUTS3).

Reports

[1] Van Delden, Windmeijer and ten Bosch, *Finding enterprise websites,* The European Establishment Statistics Workshop, Sep 24-27 2019, Bilbao, Spain, https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WPC_Experimental_statistics

[2] Van Delden, Windmeijer and ten Bosch, *Searching URLs of business websites,* Discussion paper Statistics Netherlands, expected to be released on www.cbs.nl end of 2019.