ESSnet Big Data WP2: Webscraping Enterprise Characteristics

Methodological note

The ESSnet BD WP2 performs joint web scraping experiments following in multiple countries, using as much as possible the same methodological concepts. The aim is to derive experimental statistics on enterprises from information found on the web, especially the websites of enterprises. 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 the output of research experiments and they do not necessarily align with the official statistics published on this subject.

Use case: URL Inventory of enterprises
Country: BG
Date: 2018-04-10
Authors: Kostadin Georgiev and Galya Stateva

Data sources

- The Statistical Business Register (SBR)
- Responses from Bulgarian ICT survey
- URL Inventory of enterprises from the SGA-I
- Search results from Jabse Search API, Google Custom Search API and Bing Search API on the base of the enterprise's name

Population

The enterprises as defined by the ICT survey (>10 persons, limited NACE category)

Methodology

I. Methodological procedures with BNSI software:

1. Get enterprises from the ICT population 2017 with the following data: IDs, Names, URLs, E-mails and other characteristics form the SBR in CSV file.
2. Create MySQL database table with IDs, Names, URLs, E-mails and other characteristics fields from the CSV file of the enterprises.
3. Upload the data from CSV file to the database table.
4. Get Google Search API Key and Jabse Search API Key.
5. Configure the information in the conf.php file.
6. Add necessary database table fields according to the information in the conf.php file where the output information from the project's software will be saved.
7. Upload the known and verified URLs of the enterprises from the previous project SGA-I to the database table for sustainability results.
8. Check if the known URLs from SBR have working web sites, generate URLs from the e-mails domains of the enterprises (excluding the popular e-mails domains: Yahoo, Gmail, etc. form the list in the conf.php file.) and check the generated URLs for working web sites with geturl.php script.

9. Check the known and verified URLs of the enterprises from the previous project SGA-I for working web sites with checkoldurl.php script since the changes have been likely occurred from the previous period.

10. Run google_search.php script to get up to 10 suggested URLs of the enterprises from the results of running Google Search API with the enterprises names.

11. Run jabse_search.php script to get up to 20 suggested URLs of the enterprises from the results of running Jabse Search API with the enterprises names in Bulgarian and transliterated in Latin.

12. Make a list with enterprises IDs and a corresponding list with enterprises names from MySQL database containing enterprises' data for the ICT survey 2017 target population.

13. Run ISTAT software UrlSearcher.jar either stands alone or through UrlSearcher.php script to get to up to 10 suggested URLs of the enterprises form the results of running Bing Search API with the enterprises names.

14. Run list.php script to make manual verification of the correct URLs of the enterprises from the checked URLs of the SBR and previous project SGA-I and the suggested URLs from the Google, Jabse and Bing Search APIs. After running this step, you have a List of enterprises with known URLs in the DB table.

15. Run the info.php script to see statistics from the above executed steps for URLs Retrieval at regional level and NACE categories. .

The BNSI URLs retrieval scripts are available at https://github.com/kostadingeorgiev/bnsi_bigdata

The ISTAT software UrlSearcher is available at https://github.com/SummaIstat/UrlSearcher

II. Methodological procedures with ISTAT software:

1. Make a list with enterprises IDs and a corresponding list with enterprises names from MySQL database containing enterprises' data for the ICT survey 2017 target population.

2. Run ISTAT software UrlSearcher.jar either stands alone or through UrlSearcher.php script to get to up to 10 suggested URLs of the enterprises from the results of running Bing Search API with the enterprises names (URLs Searching).

3. Make a negative list with domains of yellow pages sites.

4. Run RootJuice.jar with the negative domain list and the result seed.txt file from the execution of the UrlSeracher.jar program.

5. Get running Solr 4.10.4 and create a data collection.

6. Configure the data collection and run the SolrTSVImporter.jar with the result file from the RootJuice.jar program, to populate the Solr data collection (URLs crawling).

7. Configure the UrlScorer.jar parameters, crate a list with territorial units, create a list with enterprises IDs, Names and Address information, and run UrlScorer.jar program (URLs scoring).

8. Create a list with enterprises IDs, Names and known URLs, and run the UrlMatchTableGenerator.jar program using the result file from the execution of the UrlScorer.jar program **(Machine learning with Custom R script)** taking into account that a subset of enterprises for which the correct link is already indicated is available. Our input training dataset consists 24 268 records that had at least one page fetched. On the basis of the output scoring dataset we first associated to each enterprise the link with the highest score. As we know if the link is correct or not, a dichotomous variable correct_Yes_No says if the URL is the right one or not: this variable plays the role of the Y variable, to be predicted by the model. Together with this information, variables indicating success or failure of the search of telephone, VAT code, municipality, province and zip code play the role of the X variables (predictors), together with the link position and coincidence of the central part of the URL with the name of the enterprise (simple URL). This initial set is split into two equal size subsets, the first acting as the proper training set to fit the model, the second as the test set used to evaluate the performance of the model. Three different models were fitted: logistic, neural network and random forest. Their performance was almost equivalent, and logistic was chosen for the interpretability of its parameters ***(exactly the same as ISTAT procedure, which is well described in the Methodological note of the use case: URL retrieval).***

9. Apply the logistic model to the set of enterprises for which the website URL was not known.

The ISTAT software is available at https://github.com/SummaIstat/.

Related figures

I. The results from URLs Retrieval (with BNSI software) procedure is:

The total size of population (enterprises with 10+ employees) of the ICT survey 2017: 27489
Number of e-mails of the enterprises in the population: 19888
Initial number of URLs of enterprises in the SBR: 1994
Verified URLs from the initial ones: 1822
Verified URLs from the e-mails: 6844
Number of serches in www.jabse.com: 18245
Number of serches in www.jabse.com Latin: 19043
Number of serches in www.google.com: 27356
Number of serches in www.bing.com (with ISTAT software): 27489
URLs of enterprices found with previous project SGA-I for the target population of the ICT survey 2017: 9024
URLs of enterprices found with the project SGA-II: 11442

The benchmark analysis was carried out between the ICT survey 2017 data and URL Retrieval data (ICT survey question: *Does your enterprise have a website (Yes, No)*?). The total 27489 enterprises were object to URL Retrieval procedure and 4776 of them were in the scope of the ICT survey 2017. The number of full matches (*Yes/Yes* and *No/No* results in the both sources) are 3907 or 81.80% success of the URLs retrieval procedure.

II. The results from URLs Retrieval (with ISTAT software) procedure is:

The URLMatchTableGenerator takes the results from the URLScorer and compares them with known list of enterprise's URLs (11442 URLs). The result shows that software predicts the right URLs of 74 % of the enterprises. The results are slightly better than those from SGA-I, but there is room for improvement by adopting a better list of yellow pages and internet catalogues. Also, there were differences between the expected data fields from the software and the provided fields from the Bulgarian SBR, for example the area code and phone number of the enterprise are concatenated in Bulgarian SBR, compare to their separate use of in the ISTAT software.

Matching	Count
URLs that match	8067
URLs that don't match	2810

The results from Analyse phase with Machine learning method are the following:

Values "0.281" and "0.488" are such that all first seven classes are defined as "non-links", while the last two classes are defined as "links"; the eighth class is destined to manual inspection.

```
using threshold 8
URL to be taken
        0         1
0.8022087 0.1977913

URL to be excluded
        0         1
0.4310615 0.5689385

URL to be crowdsourced
        0         1
0.7667299 0.2332701
```

Accordingly to the predicted score, we decided if the found link was acceptable or not in terms of reliability. We were able to find 898 new URLs , which were added to the already available URLs from deterministic approach and SBR, meaning that a total number of URLs is 12 340 or 44.9% of enterprises (with 10+ employees) have web-sites. Considering that the estimate (from the ICT survey) of enterprises with a website is 50.8% the obtained coverage is 88%.

```
Logistic with 10 classes
Upper five classes taken as correct links

      score_class true false classification_error group
1   [0.0025,0.0176]   49  1171            0.9598361     1
2   (0.0176,0.0411]   58  1157            0.9522634     2
3    (0.0411,0.094]   84  1320            0.9401709     3
4     (0.094,0.139]  108  1076            0.9087838     4
5     (0.139,0.201]  273  1568            0.8517110     5
6     (0.201,0.255]  127   361            0.7397541     6
7     (0.255,0.281]  367   915            0.7137285     7
8     (0.281,0.488]  470   624            0.5703839     8
9     (0.488,0.802]  533   332            0.3838150     9
10    (0.802,0.949] 1263   278            0.1804023    10


Recall:  0.8283313
Precision:  0.5237192
F1 measure:  0.6417112
```

The first results from usage of the machine learning techniques with ISTAT software are promising in terms of quality and efficiency.

Limitations and future work

- The ICT survey was carried out in the beginning of 2017, where the URLs retrieval procedures were performed in January 2018. We intend to repeat the comparison when the ICT survey 2018 results are available. It's probably will lead to the more accurate results;
- We used SBR data, SGA-I results and three search engines for improving of the accuracy and to facilitate the manually verification of the URLs enterprises;
- The output results are better than SGA-I results and we may conclude that URLs Retrieval procedure could be used in the real statistical production and improving the quality of the SBR data.
- The ISTAT score vector is not completely relevant to the Bulgarian SBR data. The fine-tuning of the scores should be applied and the vector could be modified more precisely to the Bulgarian SBR, e.g. municipality name to be replacing with street name; telephone number format to be precise and etc.
- We could use the known URLs from SGA-I or SGA-II results in the future to retrieve the URLs only for the rest of the enterprises from the target population. It'll decrease the staff burden and the time spent.