Austrian Journal of Statistics
April 2015, Volume 44, 31–43.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i2.53

Internet as Data Source in the Istat Survey on ICT in Enterprises

Giulio Barcaroli
Istat
Alessandra Nurra
Istat
Sergio Salamone
Istat
Monica Scannapieco
Istat
Marco Scarnò
Cineca
Donato Summa
Istat

Abstract

The Istat sampling survey on *Information and Communication Technologies (ICT) in enterprises* aims at producing information in particular on the use of Internet and other networks by Italian enterprises for various purposes (e-commerce, e-skills, e-business, social media, e-government, etc.). To such a scope, data are collected by means of the traditional instrument of the questionnaire. Istat began to explore the possibility to use web scraping techniques, associated, in the estimation phase, to text and data mining algorithms, with the aim to replace traditional instruments of data collection and estimation, or to combine them in an integrated approach. The 8,687 websites, indicated by the 19,114 enterprises responding to the survey of year 2013, have been *scraped* and the acquired texts have been processed in order to try to reproduce the same information collected via questionnaire. Preliminary results are encouraging, showing in some cases a satisfactory predictive capability of fitted models (mainly those obtained by using the *Naïve Bayes* algorithm). Also the method known as *Content Analysis* has been applied, and its results compared to those obtained with classical learners. In order to improve the overall performance, different systems for web scraping and mining have been experimented and evaluated. On the basis of the final results of this test, an integrated system harnessing both survey data and data collected from Internet to produce the required estimates will be taken into consideration, based on systematic scraping of the near 100,000 websites related to the whole population of Italian enterprises with 10 persons employed and more, operating in industry and services. This new approach, based on *Internet as Data source (IaD)*, is characterized by advantages and drawbacks that need to be carefully analysed.

Keywords: web scraping, web mining, data mining, text mining, Internet as Data source, Big Data, R.

1. Introduction

Internet can be considered as a data source (belonging to the vast category of Big Data), that may be harnessed in substitution of, or in combination with, data collected by means of the traditional instruments of a statistical survey. In case of substitution, the aim is to reduce

respondent burden; in case of integration the increase in accuracy of the estimates is the main goal. The *Community survey on ICT usage and e-commerce in enterprises* (in short, *ICT in enterprises*) carried out by Istat (together with all EU Statistical Institutes) is a natural candidate to experiment this approach, as the questionnaire contains a number of questions, related to the characteristics of the websites owned or used by the enterprises, whose answers can be deduced directly by the content of these websites. An experiment has been conducted, whose aim is twofold:

- from a technological point of view, to verify the capability to access the websites indicated by the enterprises participating to the sampling survey, and collect all the relevant information;
- from a methodological point of view, to use the information collected from Internet in order to predict the characteristics of the websites not only for surveyed enterprises, but for the whole population of reference, in order to produce estimates with a higher level of accuracy.

Previous work on the usage of Internet as a data source for Official Statistics was carried out by Statistics Netherlands in recent years (ten Bosch and Windmeijer 2014). In particular, a first domain of experimentation was related to *air tickets*: the prices of air tickets were collected daily by Internet robots, developed by Statistics Netherlands supported by two external companies, and the results were stored for several months. The experiment showed that there was a common trend between the ticket prices collected by robots and existing manual collection (Hoekstra, ten Bosch, and Harteveld 2012). Two additional domains of experimentation were *Dutch property market* and *clothes prices*, the first exhibiting more regularity in the sites structure, the latter more challenging with respect to automatic classification due to lack of a standard naming of the items, and variability in the sites organization. The scraping task described in this paper goes a step further with respect to such experiences, by collecting data without any assumption on the structure of the websites and by providing the ability to scale up to a huge number of them.
This paper is organized as follows. In section 2, a general description of the survey is given with a focus on the section of the questionnaire interested to the experiment. In section 3, some different solutions for the web scraping system are described. In section 4, different inference approaches are outlined, together with their results. In the conclusions, pros and cons of the *Internet as Data source* approach are evaluated, and indications about future work are outlined.

2. Description of the survey

The *ICT in enterprises* survey is carried out annually by the Italian National Statistical Institute (Istat)[1], according to a common questionnaire and a harmonised methodology set out by Eurostat, shared in all the EU member states and in cooperation with OECD. The survey collects information on ICT usage by enterprises with 10 and more persons employed working in industry and services[2] and, in particular, involves a sample of small and medium firms and all the large enterprises (with at least 250 persons employed). The survey, on the basis also of a benchmarking framework adopted for the Information Society policy, is annually adapted to the needs of users and policy makers. Moreover, technological evolution

[1]For a complete description, see http://siqual.istat.it/SIQual/visualizza.do?id=5000078

[2]The enterprises are classified in the following economic activity (NACE Rev. 2): Manufacturing; Electricity, gas and steam, water supply, sewerage and waste management; Construction; Wholesale and retail trade repair of motor vehicles and motorcycles; Transportation and storage; Accommodation and food service activities; Information and communication; Real estate activities; Professional, scientific and technical activities; Administrative and support activities; Repair of computers. In 2013 the sample was of 32,328 enterprises and the frame population of 193,130 enterprises. The survey frame is represented by the Italian Business Register of active enterprises (BR). The sampling design is stratified with one-stage selection of units with equal probability. Strata are defined by the combination of economic activities, size classes and administrative regions of the administrative office of enterprises.

requires flexible statistical measurements of the phenomena observed and this survey responds to the need to better tailor some indicators from year to year while keeping the others fixed and more comparable in accordance with the general criteria of reduction or maintenance of response burden on enterprises within a given limit. For ICT survey this limit was fixed to 66 variables per questionnaire and was one of the main reason to begin discussing about the use of Internet as source of data and the possibility to substitute or to complete the information asked through more traditional statistical instruments like self-administered survey. The survey aims at measuring the adoption of ICT, broadband Internet connection, website functionalities, the impact of new technologies on the relationships with customers and suppliers (sharing information electronically on Supply Chain Management, exchanging automatically business documents), on organizational and marketing aspects (sharing electronically information on sales and/or purchases with any internal function, using applications to analyse information collected on clients), e-commerce, e-government. Figures considered in this paper derive from raw survey data of year 2013. In 2013 respondents to the ICT survey were 19,114, equal to 59% of the total initial sample and 9.9% of the universe of Italian active enterprises with 10 and more persons employed. The ICT questionnaire includes a section on access and use of the Internet with a subsection on use of website that is the subject of this paper. We used information coming from questions about facilities supplied through website and those given by respondents in a final section dedicated to enterprises, indicating the website URL. The observed variable (*Does your enterprise have a Website or Home Page, or one or more Internet pages?*) does not refer specifically to the ownership of the website, but to the use of a website by the enterprise to present its activities. The enterprises answering 'yes' to this filter question can include not only the existence of a website which is located on servers belonging to the enterprise, but also third party websites (e.g. one of the group of enterprises to which it belongs, other third party websites[3]). This definition represents a first possible limit of this experimental study, as we will discuss later. For the enterprises having a website, this question focuses on the measurement of its specific uses. In particular we concentrated our study on the possibility to sell product or services via web (*Online ordering or reservation or booking, shopping cart facility*, from now on *Web sales functionality*)[4]. This choice was due to particular potential importance of this facility for positive impact on enterprise's performance, for giving a measure of level of e-business readiness and intensity of firms and sectors and also because, in terms of web contents, should be easier to recognize keywords that could detect the same phenomenon through automated tools.

3. The web scraping system

Web scraping is the process of automatically collecting information from the World Wide Web, based on tools (called scapers, internet robots, bots etc.) that navigate and extract the content of a website, and store scraped data in local data bases. Web scraping may be against the terms of use of some websites: courts are prepared to protect proprietary content of commercial sites from undesirable uses, even though the degree of protection for such content is not clearly settled. The amount of information accessed and copied depends on the degree to which the access is perceived as adversely affecting the site owner's system, and the types and manner of restrictions to such access.

[3]An enterprise may offer web sales functionality and still not have a website as the sales are through e-marketplaces that are not included in questions considered in this paper.

[4]From the Methodological Manual (2013): 'This item refers to a facility which allows the user to order products or services with no additional contact offline or via e-mail necessary (for the ordering). It includes also websites which allow the reservation of hotel rooms or the booking of flights. It does not include a link in the website which directs the user to an e-mail application which requires the user to send the order via e-mail. Payment may or may not be included in the ordering facility, e.g. payment may be made on reception of the product also by other means than electronic payment'. Inside the benchmarking framework 2011-2015 it is included the indicator D7 asking for percentage of enterprises having a website with web sales facilities (Website or a Home Page with online ordering or reservation or booking).

In the following, different solutions for the web scraping are described: the first one is already available and has been used specifically for this experiment, while the others are still under investigation. Indeed, we are carrying out a dedicated activity with the purpose of testing and comparing different technological solutions for scraping in order to figure out the most suitable solution for a specific purpose.

3.1. The web scraping application based on JSOUP and ADaMSoft

A first choice was to develop the scraping application by referring to an open source library called JSOUP (available at `http://jsoup.org/`) and by integrating it in the ADaMSoft system (available at `http://adamsoft.sourceforge.net/`). JSOUP is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM (Document Object Model), CSS (Cascading Style Sheets), and jQuery-like methods. More in detail this tool allows:

- scraping and parsing HTML from a URL, file, or string;
- findind and extracting data, using DOM traversal or CSS selectors;
- manipulating the HTML elements, attributes, and text.

On the other side, ADaMSoft was selected because it included facilities in managing huge data sets and because it already contains procedures to deal with textual information. We then developed an appropriate step that can be considered a mix between a web spider (i.e. a tool to extract the structure of a website) and a retriever for both the content and the tags of identified pages. More specifically, the content is intended as the text that can be viewed when browsing a page, while the tags are all the hidden HTML keywords that guide the way in which the page is displayed, the actions associated to a button, etc.

Inside such step we implemented methods that permit to:

- keep into account the limitations eventually defined in the `robots.txt` file;
- extract the structure of the website for a given level of depth (i.e. the sub-links from the main URL);
- filter the resultant URLs (in order to avoid, for example, those links that redirect to other websites);
- emulate a specified user agent;
- pass to the website a series of cookies;
- specify the method (GET or POST);
- use a given time limit to explore the website;
- identify and access different content types (HTML, obviously, but also PDF, DOC, etc.).

The input to the procedure is a dataset containing the identifiers of the enterprises, the indication of related URLs and also the indication of the level of depth that is considered as acceptable in the exploration of the websites (i.e. to what extent sub-links will be taken into consideration: in our tests we set this parameter to '2' and '3'). To increase the efficiency of the task, the procedure permits to examine at the same time more than one website.

We considered really crucial to retrieve also the elementary information of a HTML tag (i.e. its type, name and content), because it could contain discriminant terms that can help us in identifying the nature of the website; for example a button associated to an image called "paypal.jpg" could be a clear sign of web sales functionality.

Running the Web Scraper procedure for the original 8,687 websites took less than one day on a Windows PC platform. Actually we observed some difficulties in accessing some websites; these derive from a not correct specification of the main URL, and/or making use of technologies not entirely based on standard HTML text (like, for example, the websites realized with Flash technology).

By considering only those websites for which at least a page was accessed, we found an average value of 235,108 characters retrieved.

For what concerns the HTML tags (that we restricted to one of the following types: *address, button, fb:like, form, label, menu, input, meta, option, rss, select, textarea*), we collected more than 17.5 million of elementary information contained in tags (which corresponds to an average of 2,649 tags collected for each one of the 6,632 websites for which at least one tag was retrieved).

Due to the huge amount of terms, we proceeded by tokenizing each of these by transforming all non valid ASCII code characters in spaces (i.e. `paypal.jpg` is transformed in two terms: `paypal` and `jpg`) and we considered their main lemma, by deleting all the determiners, the articles, the prepositions, etc. (i.e. all the terms that can be considered generic). To this purpose, we used a package named TreeTagger, directly executed from inside Adamsoft.

TreeTagger is a tool for annotating text with part-of-speech and lemma information, developed at the Institute for Computational Linguistics of the University of Stuttgart (`http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`). It has been used by referring to the Italian and to the English lexicon in order to permit the selection of less than 60,000 terms to be considered as the basis for further processing steps.

3.2. Other solutions for web scraping: Nutch/Solr suite and HTTrack

The Apache suite used for crawling, content extraction, indexing and searching results is composed by Nutch and Solr. Nutch (available at `https://nutch.apache.org/`) is a highly extensible and scalable open source web crawler, it facilitates parsing, indexing, creating a search engine, customizing search according to needs, scalability, robustness, and scoring filter for custom implementations. Built on top of Apache Lucene and based on Apache Hadoop, Nutch can be deployed on a single machine as well as on a cluster, if large scale web crawling is required. Apache Solr (available at `https://lucene.apache.org/solr/`) is an open source enterprise search platform that is built on top of Apache Lucene. It can be used for searching any type of data; in this context, however, it is specifically used to search web pages. Its major features include full-text search, hit highlighting, faceted search, dynamic clustering, database integration, and rich document handling. Providing distributed search and index replication, Solr is highly scalable. Both Nutch and Solr have an extensive plugin architecture useful when advanced customization is required.

Starting from a list of URLs (root pages), Nutch fetches, parses and indexes for each of them all the linked resources according to a series of constraints, the most important are (i) the link depth from the root page that should be crawled, (ii) the maximum number of pages that will be retrieved at each level up to the depth. Nutch offers a series of fine configurable NLP (Natural Language Processing) functions applicable on fetched web resources, such as tokenization, stop-words removal and stemming. Finally Nutch delegates searching to Solr.

All problems encountered are relative to environment configuration and tools integration, for example it was necessary to manage case sensitivity, site load balancing, page redirections, plugins and OS configuration, etc..

Although this web scraping approach requires an initial effort in terms of technological expertise, in the long run it can lead to a substantial return on investment as it can be used on many other contexts to access Big Data sources. As an example, it can be used as a platform to access and analyse web resources like blogs or social media to perform semantic analyses or sentiment analyses tasks.

HTTrack (available at `http://www.httrack.com/`) is a software tool that permits to "mirror" locally a web site, by downloading each page that compose its structure. For the specific case study described in this paper, we access URLs of enterprises websites and download related HTML resources. Once such resources are locally available, we are able to access specific the content of HTML elements (e.g. title, HTML links, body, etc.) that are used for subsequent analysis steps.

The main differences between HTTrack and Nutch/Solr are:

- generic parsing and indexing tasks are only performed by Nutch/Solr;
- direct processing of HTML resources is easily available with HTTrack, while it requires dedicated effort with Nutch/Solr/Lucene.

According to our first tests, HTTrack approach results to be more suitable for the scraping task of the case study here described, as it requires access to specific HTML pieces.

4. The inference system

Once completed the web scraping activities, before proceeding with the inference phase a pre-processing step was applied, consisting in treating the text terms (reduction to lower case, elimination of punctuation and stop-words, stemming) and in selecting only the words that showed a significant influence on the target variables. This influence was determined by a two-step procedure:

- a first selection was made by applying a *correspondence analysis* bewteen a given target variable and the words contained in the scraped texts;
- a second selection was obtained by evaluating the chi-square associated to the cross-classification of a given target variable with respect to the presence/absence of a given word.

By applying the first step a subset of words can be selected, still too numerous to be managed in the modelling phase. This is why we applied the second step, in which four different subsets of words have been defined: having set as thresholds the percentiles 99.5, 99.0, 97.5, 95.0 related to their chi-square distributions, only words with a chi-square exceeding those thresholds have been considered.

The final result of this pre-processing consists in a document/term matrix, where each row represents a website, each column is referred to an influent word, and the intersection indicates the frequency (or just the presence or the absence) of the word in the website. In order to choose the best instruments useful to build the inference system, in this exploratory phase we tested several of them, distinguished in:

- data mining learners, applicable to this text mining problem: *Classification Trees*, ensemble learners (*Random Forest, Adaptive Boosting, Bootstrap Aggregating*), *Neural Networks, Maximum Entropy, Support Vector Machines, Latent Dirichlet Allocation* (James, Witten, Hastie, and Tibshirani 2013);
- the approach followed in the *Content Analysis* (Hopkins and King 2010);
- the learner most suitable for text mining: *Naïve Bayes* (Lantz 2013).

As usual, available data have been partitioned in a training set and in a test set: each model, fitted using the training set, has been applied to the test set in order to evaluate its performance, by comparing observed and predicted values for the target variables, both at individual and aggregate level. In general, the proportion between the two sets was determined in 75/25, but a sensitivity analysis has been performed for Naïve Bayes and Content Analysis defining 9 different rates for the training set (from 0.1 to 0.9). Experiments have been carried out considering the four different subsets of words defined accordingly to their chi-square, and the most favorable in terms of performance has been retained.

Performance has been measured by considering the following indicators: (i) *precision rate* (number of correctly classified cases on the total number of cases), (ii) *sensitivity* (rate of correctly classified positive cases), (iii) *specificity* (rate of correctly classified negative cases). In addition, we also introduced (iv) the *proportion of predicted positive cases*, as it corresponds to the final estimate that we want to produce, and whose accuracy we want to maximize. These four indicators can be easily computed from the *confusion matrix*.

4.1. Data mining learners

In Table 1 we report the results of the application of the different learners in order to predict web sales functionality (we made use of R packages **RTextTools** (Jurka, Collingwood, Boydstun, Grossman, and Atteveldt 2014) and **rattle** (Williams 2011)) (R Core Team 2014). It is possible to notice that the precision level is in general acceptable: it ranges from a minimum of 79% to a maximum of 85%. Specificity is always very high. The real problem is given by sensitivity, that is the capability to correctly classify positive cases, i.e. the websites that offer web sales functionality: in many cases its value is too low to be considered as acceptable. As for the proportion of web sales functionality, in general data mining learners fail in reproducing the correct aggregates.

Table 1: Performance indicators for data mining learners (variable *Web sales functionality*).

METHOD	**Precision**	**Sensitivity**	**Specificity**	**Proportion Web sales = Yes (observed)**	**Proportion Web sales = Yes (predicted)**
Classification tree	0.83	0.28	0.98	0.21	0.08
Random forest	0.85	0.34	0.99	0.22	0.08
Bootstrap aggregation	0.82	0.48	0.91	0.21	0.10
Adaptive boosting	0.80	0.39	0.91	0.22	0.17
Maximum entropy	0.80	0.46	0.90	0.22	0.18
Support Vector Machines	0.79	0.02	0.99	0.22	0.01
Neural networks	0.82	0.21	0.98	0.20	0.06
Latent Dirichlet Allocation	0.81	0.18	0.98	0.21	0.05

4.2. Text mining specific approaches

Content analysis

Hopkins and King (2010) proposed a method quite different from all the others so far considered, as it does not require statistical or machine learning modeling of data and consequent individual predictions. It does not even require a training set to be a representative sample of the whole population: the only requirement is that the training set must contain a sufficient number of cases for each combination of terms.

In order to verify the robustness of this method, different training sets have been obtained by drawing samples from the available websites, varying the sampling rate from 0.1 to 0.9 (100 samples for each sampling rate), and related estimates of *Web sales functionality* rate have been produced for each sample by using Content Analysis. The software used for its application is described in Hopkins, King, Knowles, and Melendez (2012), and is available at `http://gking.harvard.edu/readme`.

It can be seen (Figure 1) that, especially in cases from 0.1 to 0.3 of training set rate, the method seems to be unbiased, as the mean of the estimates tends to coincide with the proportion calculated in the total number of cases. But the range of the estimates is considerably large: for example, in the case of 0.1, interval of estimates goes from 0.08 to 0.31, and we can observe even worse situations for the other training set rates.

Naïve Bayes algorithm

The Naïve Bayes algorithm is the most used in the field of the text classification, where it can be considered as a standard choice. It is called "naïve" because of its (simplistic) assumptions concerning data, as it assumes that all the features in a dataset are independent and equally important, a condition that is seldom verified in real situations. Actually, words in a text are not equally important in order to predict a given category to be associated to the text, and words are not independent each other. But Naïve Bayes works well despite the fact that its basic assumptions are very seldom fulfilled. We made use of the implementation available in the R package **e1071** (Meyer, Dimitriadou, Hornik, Weingessel, and Leisch 2014). In Table 2 the results obtained by the application of Naïve Bayes are reported.

Table 2: Confusion matrix for Naïve Bayes application (variable *Web sales functionality*).

	Predicted Values			
Observed Values	**1 (YES)**	**2 (NO)**	**Total**	**Relative Frequencies**
1 (YES)	120	119	239	0.22
2 (NO)	121	748	869	0.78
Total	241	867	1,108	1.00
Relative Frequencies	0.22	0.78	1.00	

From this confusion matrix it is possible to calculate the usual performance indicators (Table 3). It can be seen that Naïve Bayes is slightly inferior to some data mining learners in terms of precision, but performs better in terms of sensitivity, and reaches a practically perfect coincidence between the predicted proportion and the observed one.

Table 3: Values of performance indicators for Naïve Bayes application (variable *Web sales functionality*).

Indicator Name	**Indicator Value**
Precision	0.78
Sensitivity	0.50
Specificity	0.86

As in the case of Content Analysis, also for evaluating the robustness of Naïve Bayes solutions a simulation has been carried out, under the same setting.

The graph in Figure 2 shows that the method is slightly biased[5], as it systematically overestimates the true value (in the order of one or two percentage points). But the variability of the estimates is much lower than in the case of the Content Analysis: considering the case related to the training set rate equal to 0.1, the range goes from 0.19 to 0.24.

[5]In presence of bias, a method to correct the aggregations resulting from individual predictions obtained by a given learner has been proposed by Hopkins and King (2010). Given a variable D with two possible values (1 and 2), we know that

$$P(\hat{D} = 1|D = 1) : sensitivity \tag{1}$$

$$P(\hat{D} = 2|D = 2) : specificity \tag{2}$$

Then, by the law of total probability:

$$P(\hat{D} = 1) = (sensitivity)P(D = 1) + (1 - specificity)P(D = 2) \tag{3}$$

we can obtain:

$$P(D = 1) = \frac{P(\hat{D} = 1) - (1 - specificity)}{sensitivity - (1 - specificity)} \tag{4}$$

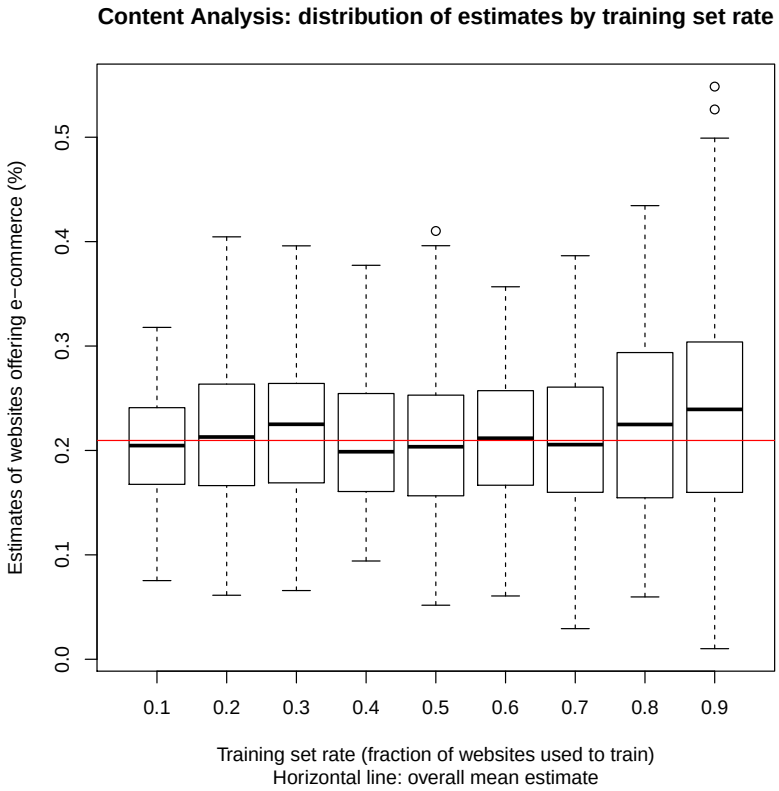

Figure 1: Content Analysis: distributions of estimates calculated on test sets varying the training set rate (variable *Web sales functionality*)

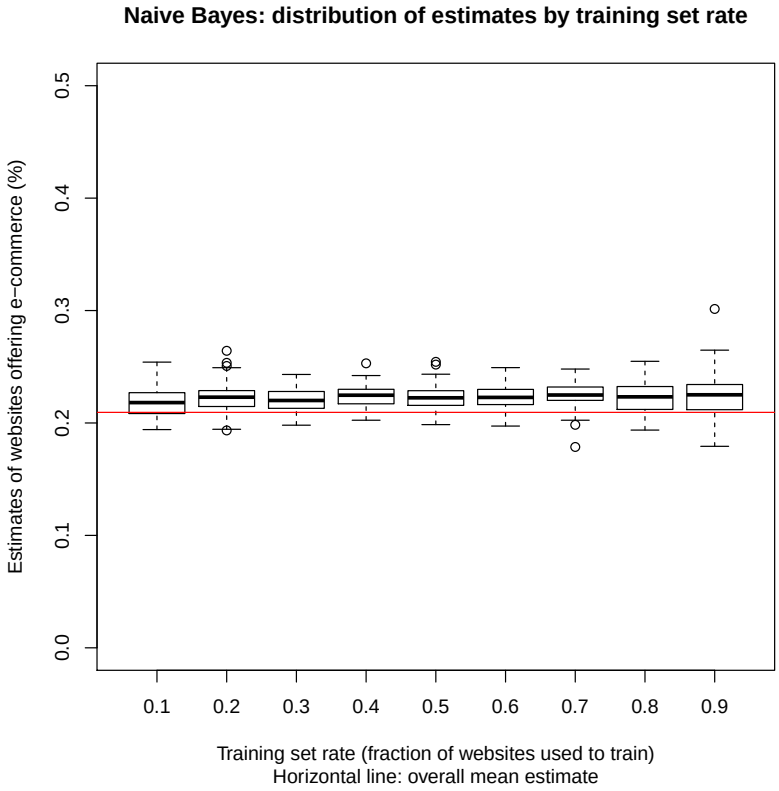

Figure 2: Naïve Bayes: distributions of estimates calculated on test sets varying the training set rate (variable *Web sales functionality*)

As it resulted to be the best method among those considered, Naïve Bayes has been applied to other suitable variables in the questionnaire, obtaining the results reported in Table 4.

As for the software that has been used to produce the results reported above, the ADaMSoft modules for web scraping and texts handling, together with the R scripts for the application of Naïve Bayes algorithm, are available at http://adamsoft.sourceforge.net/appscripts.html.

Table 4: Results of the application of Naïve Bayes to the complete set of B8 question.

QUESTION	Precision	Sensitivity	Specificity	Proportion Web sales = Yes (observed)	Proportion Web sales = Yes (predicted)
Web sales functionality	0.78	0.50	0.86	0.21	0.21
Orders tracking	0.82	0.49	0.85	0.18	0.11
Description and price list of goods	0.62	0.44	0.79	0.48	0.32
Personalised content for regular visitors	0.74	0.41	0.781	0.09	0.23
Possibility to customise online goods	0.86	0.53	0.87	0.05	0.14
Privacy policy statement	0.59	0.57	0.64	0.68	0.51
Online job application	0.69	0.521	0.78	0.35	0.33

5. Conclusions

The best method resulting from the experiment seems to be the Naïve Bayes. The values of the first three indicators of performance (precision, sensitivity and specificity) are all good, and the forth (alignment between observed and predicted aggregate) is the best with respect to the other learners. It is slightly biased with respect to the Content Analysis, but is much better in terms of variability of the estimates.

With regard to the relatively low levels of sensitivity, due to the high number of false negatives (represented by enterprises declaring in the survey to have web ordering facility but resulting as *not having* this possibility on the site-centric scraping basis), it is important to underline that the use of *website centric measurements* allows only a partial measurement of the phenomenon detected by the survey. In fact, in the questionnaire the wording of the questions permits to the respondent to answer "yes" with reference not only to the owned website but also to those sites of the linked companies (subsidiaries or owning the brand, or other third parties). Moreover, the positive answers in the survey consider also e-sales between enterprises: commercial transactions between the responding enterprise and other enterprises, named business-to-business (B2B, e.g. manufacturer and a wholesaler, a wholesaler and a retailer). With respect to business-to-consumer (B2C) e-sales or reservation systems, B2B is often based on a protected access requiring a login and a password, making difficult to identify automatically e-sales functionalities of investigated websites. These two factors can explain why using different instruments (survey vs scraping) we measure the same phenomenon but delimited by different boundaries.

The extension of the *IaD* methods to further technical indicators (i.e. number of pages, downloading speed, technical or language accessibility, etc.) requires to consider also other issues. In the following, we report the main trade-offs, strengths and weaknesses of web scraping and mining methods presented above, compared with the traditional statistical survey.

Benefits and opportunities

- in terms of accuracy: it is possible to extend the analysis to the whole population and not only a subsample (avoiding sampling errors), therefore producing more detailed figures (e.g. for enterprises with less than 10 persons employed not observed in ICT survey); degree of closeness of estimates to the true values could be improved thanks to technology and programming new code (reducing measurement errors);
- in terms of relevance of information: it is possible to discover new services, new information; to investigate other web functionalities as e.g. advertisement of open job positions or online job application, usage of website safety certificate, possibility for customer to submit electronic complaints (via e-mail, web form, etc.), links or references to the enterprise's social media profiles, etc.;
- in terms of comparability among countries: it could be improved if same automatic website centric tools are used;
- in terms of transparency of process: it is avoided human misunderstanding among concept/definition and scope of the question of survey;
- in terms of statistical burden: the respondent burden can be reduced (but we discussed about only one variable out of 66);
- in terms of timeliness: it is improved;
- in terms of reiteration of process: it is possible to repeat the entire automatic data collection during the same period of traditional survey.

Costs and disadvantages

- in terms of accuracy: it is necessary to manage and maintain a list of URLs for the entire population; there is a non-negligible risk to introduce bias into the estimates;
- in terms of coherence of measured concepts: web mining applications described may not catch the same phenomenon of ICT survey;
- in terms of comparability among countries: using different tools (survey vs IaD) or a different list of words could produce less comparability;
- in terms of technology used: there are technical limits to solve as the long run time necessary for the crawler to get the entire content; security barriers inside the website preventing automatic access (restrictions); website not in HTML (i.e. in Flash), redirect problems;
- semantic limits of automatic tools: not all services offered on websites can be well semantically delimited;
- time spent in analysis and programming: to discover new information requires to analyse data collection in different ways and then to update program code;
- in terms of development and maintenance efforts of the web mining applications: persons with high level skill are required.

The web mining (or Internet as Data source) approach experimented in the *ICT in enterprises* survey revealed to be promising and can be continued and extended in different directions:

- with reference to the population of interest: we can consider the URLs of all the units belonging to the Business Register, and perform a mass scraping of related websites (in this case also experimenting more properly the high volume problems (scaling) related to Big Data), considering the whole survey sample as a training set, so to obtain a model that can be applied the whole population. The aim is twofold: (i) to produce estimates under a full predictive approach, reducing the sampling errors at the cost of introducing additional bias (both components of Mean Squared Error should be evaluated); (ii) to identify the subpopulation of enterprises active in web sales transactions with individuals as the end consumer (B2C), that can be considered as a new sampling frame to consider in the ICT survey or useful to carry out new *ad hoc* surveys;

- with reference to the content of the questionnaire: the approach used with the set of variables contained in the 'B8' section of the questionnaire will be evaluated also with regard to other suitable sets of variables in the questionnaire (e-recruitment, use of social networks, etc.).

Anyway, it is necessary to improve the results of the web mining applications by investigating specific situations. While conceptual reasons justify, as discussed above, the high percentage of false negatives, it is more difficult to understand cases in which web scraping finds web sales functionality signals in websites contrary to answers of the survey (false positives). In the future it is necessary to better explore these false positives because, for example, they could be a signal that respondents do not understand correctly the question. Different explanations could be found in time lag between survey and web scraping, or in flaws in the methods and tools used for web scraping and text mining.

References

Hoekstra R, ten Bosch O, Harteveld F (2012). "Automated Data Collection from Web Sources for Official Statistics: First Experiences." *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics*, **28**(3-4), 99–111.

Hopkins D, King G (2010). "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science*, **54**(1), 229–247.

Hopkins D, King G, Knowles M, Melendez S (2012). *ReadMe: Software for Automated Content Analysis*. Version 0.99835, URL http://gking.harvard.edu/files/gking/files/readme.pdf.

James G, Witten D, Hastie T, Tibshirani R (2013). *An Introduction to Statistical Learning with Applications in R*. Springer Texts in Statistics.

Jurka T, Collingwood L, Boydstun A, Grossman E, Atteveldt vM (2014). *RTextTools: AutomaticText Classification via Supervised Learning*. R package version 1.4.2., URL http://CRAN.R-project.org/package=RTextTools.

Lantz B (2013). *Machine Learning with R*. Packt Publishing Ltd.

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2014). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-3, URL http://CRAN.R-project.org/package=e1071.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

ten Bosch O, Windmeijer D (2014). "On the Use of Internet Robots for Official Statistics." In *MSIS-2014*. URL http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2014/Topic_3_NL.pdf.

Williams G (2011). *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery (Use R!)*. Springer.

Affiliation:

Giulio Barcaroli
Istituto Nazionale di Statistica (Istat)
Via Cesare Balbo 16
00184 Roma, Italy
E-mail: giulio.barcaroli@istat.it

Alessandra Nurra
Istituto Nazionale di Statistica (Istat)
Via Tuscolana 1788
00173 Roma, Italy
E-mail: alessandra.nurra@istat.it

Sergio Salamone
Istituto Nazionale di Statistica (Istat)
Via Tuscolana 1788
00173 Roma, Italy
E-mail: sergio.salamone@istat.it

Monica Scannapieco
Istituto Nazionale di Statistica (Istat)
Via Cesare Balbo 16
00184 Roma, Italy
E-mail: monica.scannapieco@istat.it

Marco Scarnò
Cineca
Via dei Tizi 6/B
00185 Roma, Italy
E-mail: m.scarno@cineca.it

Donato Summa
Istituto Nazionale di Statistica (Istat)
Via Cesare Balbo 16
00184 Roma, Italy
E-mail: donato.summa@istat.it

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 44
April 2015

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2014-07-07
Accepted: 2014-10-31