

# Web scraping and web mining: new tools for Official Statistics

## *L'uso delle informazioni sul Web nelle Statistiche Ufficiali*

Sergio Salamone\*, Monica Scannapieco\*, Marco Scarnò\*\*

**Abstract** The amount of information that can be accessed through Internet is an enormous resource, useful to describe many aspects of the present society. The possibility to extract such an information has several benefits, including a “constant measurement” of a phenomenon, or the replacement (total or most probably partial) of some surveys. However, the nature of the information presents new challenges, because it is huge, sparse and unstructured. In this paper we will illustrate a study, made jointly by ISTAT and CINECA about the possibility to use web scraping techniques, associated with text mining algorithms, in order to replace traditional tools of data collection and estimation, and/or to combine them into an integrated approach. In particular, we referred to the ISTAT sampling “Survey on ICT Usage and e-Commerce in Enterprises” and to its section in which it is treated the use of Internet by Italian enterprises for e-commerce. 8,600 websites have been “scraped” and acquired texts have been processed in order to attempt the collection of the same information gathered via questionnaire. Preliminary results are quite encouraging, stating an almost satisfactory predictive capability of the fitted models.

**Abstract** *L'informazione che può essere derivata dal Web costituisce un'enorme risorsa utile a descrivere molti aspetti della Società attuale. La possibilità di estrarre questa informazione ha molti meriti; tra questi il “monitoraggio costante” di un fenomeno, o il rimpiazzare (totalmente o più probabilmente parzialmente) specifiche indagini. Tuttavia la natura dell'Informazione presenta tuttavia nuove sfide, che derivano dal suo essere enorme, sparsa e non strutturata. In questo lavoro illustreremo un'attività svolta congiuntamente dall'ISTAT e dal CINECA*

---

<sup>1</sup>

\*Sergio Salamone, ISTAT; email: [sesalamo@istat.it](mailto:sesalamo@istat.it)

\*Monica Scannapieco, ISTAT; email: [scannapi@istat.it](mailto:scannapi@istat.it)

\*\*Marco Scarnò, CINECA; email: [m.scarno@cineca.it](mailto:m.scarno@cineca.it)

*riguardante la possibilità di usare tecniche di “web scraping”, associate ad algoritmi di text mining, al fine di sostituire alcuni strumenti tradizionali di raccolta e stima dei dati. In particolare ci siamo riferiti all’Indagine Campionaria ISTAT sull’“Uso dell’ICT e dell’e-Commerce nelle Imprese” e della sezione del questionario nella quale è trattato l’uso d’Internet dalle Imprese italiane per scopi di e-Commerce. 8,600 siti web sono stati analizzati e i testi raccolti processati al fine di riprodurre le stesse informazioni collezionate mediante i questionari. I risultati preliminari sono soddisfacenti ed evidenziano una buona capacità predittiva dei metodi utilizzati.*

**Key words:** text mining, web scraping, internet as data source, correspondence analysis, word stemming

## 1 Introduction

The High Level Group for the Modernisation of Statistical Production and Services recognized that the abundance of data available nowadays should trigger a profound rethink of products offered by National Statistics Institutes (NSIs), and highlighted the risk for NSIs to be superseded by non-official statistics providers that are readier to exploit the opportunities offered by Big Data [1].

NSIs started investigating the roles that Big Data can have in Official Statistics (OS) either for use on its own, or in combination with more traditional data sources such as sample surveys and administrative registers [2]. Recently, the Scheveningen memorandum [3], which has the role of providing strategic guidelines to European national offices, clearly stated that, given the opportunities that Big Data offer to OS, NSIs are encouraged to undertake initiatives to examine the potential of Big Data sources in that regard.

Among the different possible types of Big Data sources, Internet data are surely among the most at hand and promising; Internet As a Data source (IaD) has been more and more emerging as a paradigm that concretely allows to complement or substitute traditional statistical sources that, for official statistics, are either resulting from surveys’ questionnaires or from administrative sources.

As described in [4], there are three major types of IaD methods: (i) User-centric, (ii) Network-centric and (iii) Site-centric.

User-centric methods are based on client-side applications that monitor the client behaviour. An example is a Browser Monitor that records the web pages and the sites accessed by the user. This method could be useful for statistical offices from an economic point of view, though privacy issues must be carefully taken into account.

Network-centric methods are based on the monitoring of the network and on the measurement of some properties like the traffic type (by considering the header of the packets or the ports that the communications uses). However, there are also

instruments that look inside the content of packets, not just the header (Deep Packet Inspection). Legal and privacy issues have a significant role for these methods too.

Site-centric methods are the most mature methods and include the usage of crawlers and scrapers. Data crawling is the process of automatically collecting information from the World Wide Web. A scraper takes Web resources (documents, images, etc.), and extracts information from those resources, finalized for subsequent elaboration purposes.

In this paper, we describe an experimental project adopting IaD by using site-centric methods for collecting data for the “Survey on ICT Usage and e-Commerce in Enterprises”; the provided contribution can be summarized as follows:

- We applied a complete process covering from the collection phase of the information to the estimation of classification models;
- Though there are other examples in OS of the experimental usage of scraping techniques, our experience does successfully prove the usage of such techniques in a tricky context. Indeed, an example of previous experience is related to the scraping of product prices (see e.g. [5]), which is a task that involves some technological challenges but does not require a major semantic extraction activity on scraped data. Instead, we reached the goal of understanding much more semantic rich content, such as if an enterprises does or not web sales;
- We experimented several text mining techniques in order to estimate the services offered by an enterprise through its web sites. Though the used techniques are part of state of the art on text mining, their usage in an OS context is new.

## 2 Application Domain: Survey on ICT Usage and e-Commerce in Enterprises

The Italian National Statistical Institute (ISTAT) carries out the survey on ICT usage and e-Commerce in enterprises (ICT) with 10 and more employers working in industry annually, referring to an harmonised methodology set out by Eurostat and shared by all the EU Member States.

The statistical units in the ICT survey are selected through a sample of small and medium enterprises (those with less than 250 employers), and all the large firms. In 2013, the sample was composed by 32.328 enterprises representative of the universe of firms (193.130). The collected data are mainly qualitative and provide a wide variety of topics covering among others the following areas:

- Access to and use of computers and Internet;
- E-commerce and e-business (electronic exchange of data, sale and purchase of goods or service via computers networks).

The object of the survey is annually adapted to the needs of policy makers and users, by taking into account the rapid changes in the Information Society that

implies a flexible statistical approach. This by following the suggestions of the "Digital agenda for Europe" that represents an important item of the Europe 2020 strategy.

Among the information collected with the questionnaire in 2013, the variable considered suitable for the purposes of this experimental study refers to those enterprises that stated to have a web site that provides the possibility of an online ordering, reservation or booking (question B8A). Obviously, the URL of the main web site of each enterprise is collected from the same questionnaire.

On the 19,104 respondents to the ICT survey, 13,920 were the enterprises that confirmed to have a web site; of these, 10,264 were the URLs effectively collected. 2631 enterprises answered *yes* to the question B8A, as in table 1.

**Table 1:** Respondents to the ICT survey (2013) according to their possess of a web site and to the answers to B8a

| <b>Question</b>                            | <b>Yes</b>       | <b>No</b>        | <b>Missing</b> |
|--------------------------------------------|------------------|------------------|----------------|
| Enterprises that have a Website            | 13920<br>(74.2%) | 4845<br>(25.8%)  | 339            |
| Online ordering or reservation or booking: | 2631<br>(18.6%)  | 11484<br>(82.4%) | 4989           |

### 3 Overall Approach

From a statistical point of view, we have to find those variables (i.e. some characteristics contained in a web site) that permit to identify if an enterprise offers a given service through the web. Such a model has to be estimated on a sample of enterprise' web sites, while its discriminant capacity has to be verified on those of other enterprises. So, we had to characterize each web site by transforming its content in a series of explicative variables ( $X$ ) to be used for the purpose of estimating a relation like:  $service=f(X)$ .

To the scope, we considered the whole web site of each enterprise (i.e. all its web pages) as a single *document* characterized by a series of *terms*; the idea is then the identification of those terms that are discriminant for the classification purposes.

It has to be noticed that the *terms* does not identify just the *words* but, also, other characteristics of each web page, like the *tags* that it contains. In particular, a tag is an *html* element that has a *meaning*, intended in a semantic sense or behavioural for the browser, like: *put a button in the page (with a given image and this tooltip) that, if pressed, will cause this action*.

The steps of our work were related to:

- *Web scraping*: transform the information in each web site (unstructured) in structured data that can be stored and analysed;
- *Terms extraction and normalization*: identification of those *terms* discriminant for a given service offered by an enterprise;
- *Modelling*: estimation of some supervised classification function;

- *Assessment*: evaluation of the above functions on a set of enterprise' web sites for which it is known if they offer the related service (but that were not used to estimate the functions).

### 3.1 *Web scraping*

Web scraping is the process of automatically collecting information from the World Wide Web. It is a field with active developments sharing a common goal with the semantic web vision; it is based on a tool that navigates and extracts the content of a web site and that, further, stores the scraped data in a local Data Base.

Web scraping may be against the terms of use of some websites; however, the degree of protection for this content is not settled, and depends on the type of accesses made by the scraper, the amount of information accessed and copied, etc.

There are some software that can be used directly as web scrapers or more general-purpose libraries that can be adopted for such a purpose; because our needs were to extract as more information as possible, we chose to develop our scraper with an Open Source library called JSOUP [6]. We incorporated the routine in the package ADaMSoft [7], in order to manage both the navigation of many web sites and all the data storage requirements. We encountered some problems during this step; these were mainly due to web sites not fully reachable and to those implemented by technologies that do not permit to view their pages as standard html text (like, for example, the web sites realized with ADOBE *flash*). After the execution of this step, we obtained two data sets: one containing all the text (visible by a generic user) of each web site, the other one including the characteristics of each tag (type, name and value) contained in the html pages that were visited.

### 3.2 *Terms extraction and normalization*

In this step, we joined the text and the characteristics of the tags contained in the two data sets in order to reconstruct the terms that belong to each web site. To this purpose, we transformed each not ASCII character (for example the “/”, “&”, etc.) in spaces; then we tokenized the stream of the resulting text in order to obtain the meaningful elements (words). It has to be observed that this process permitted, for example, to refer to the name of an image used to identify a button in a web page; in fact, a value like “/basket.jpg” become “basket.jpg”.

Due to the huge amount of different terms we encountered (several millions), we processed each of them by means of a specific procedure called stemming (i.e. a process for reducing inflected words to their stem, base or root form); to this purpose, we used a software named *tree-tragger* [8], which can be directly executed from the package ADaMSoft. At the end of this step, each web site was represented

by a list of words (also repeated, because we did not reduce them by considering their unicity).

### 3.3 Modeling

Before estimating the relationship between the terms and the service offered through the web site by the enterprises, we reduced the still huge amount by referring to a geometric approach based on the Correspondence Analysis (CA) [9].

To this purpose we extracted a balanced sample of web sites (according to the answer to the question B8A) and we built the corresponding term-document matrix; in particular we considered just two documents; one characterized by those web sites referring to the answer *yes*, the other by those with *no*. We then evaluated the coordinates of the projection of each term in the first, and only, correspondence axe, and we extract only those that resulted more representative for it. We considered as the minimum contribution to the axis the value 0.00001. This permitted us to reduce the terms and to order the remaining according to their contribution.

To verify the relationship between the selected terms and the service offered by the enterprises, we then considered two different approaches:

- Each of the remaining web sites (*test set*) was considered as a supplementary document; it was projected in the correspondence axe and we evaluated its relative distance between the virtual documents representing the answer *yes* or *no* to the question B8A. In this way we were able to automatically classify the web site in one of the two classes;
- We built the whole document-term matrix, where the documents were all the web sites and the terms were the significant ones as identified by CA. In each cell of this matrix, there is the number of times a term appears in the corresponding web site. Then we estimated the parameters of some non-parametric classification functions (on the same *train set* defined before):
  - Naïve-bayes;
  - Classification Trees (recursive partitioning);
  - Random Forests;
  - Adaptive Boosting (AdaBoost);
  - Bootstrap aggregating (Bagging).

### 3.4 Assessment

To evaluate the capacity of the methods to recognize if an enterprise offers, through its web site, the possibility to *online ordering or reservation or booking*, we evaluated the predicted value for each approach/function (in relation to the web sites belonging to the *test set*). Then we built the so-called confusion matrix in order to

obtain the standard indicators for classifications problems: the *success rate*, the *sensitivity* and the *specificity*.

## 4 Results

From the original data, we observed that the distribution of the answers to *online ordering or reservation or booking* (B8A) is: *yes* (17.7%) and *no* (79.5%). Our package collected the data from 6260 web sites (72%); the remaining (28%) were: not reachable (1376, 15%) or written in a not compatible html language (1042, 13%). It is interesting to observe that we collected more than 7 millions of records regarding the tags of all the web sites (and we limited their types to just the following: *address; button; fb:like; form; label; menu; input; meta; option; rss; select; textarea*).

We then proceed in stemming and normalizing all the text collected for each web site; we further reduced the web sites by considering only those having a sufficient number of terms (minimum 1000 characters) and related to those enterprises that do not have a missing value to the question B8A. This led us to consider 4441 web sites, for which we have that 966 (21.7%) *yes* and 3475 (78.4%) *no*. We identified 58428 different terms (with a mean of 7913 terms for web site).

Correspondence analysis helps us to identify the words more related to the *yes* or *no* answers; to this purpose, we extract a balanced sample of 1400 enterprises and we apply the method to all their terms. As introduced in the methodological section, we then consider the contribution to the correspondence axe in order to identify only those representative words. A selection of these, for the *No* or *Yes* answers to the *online ordering or reservation or booking* are, respectively, in fig. 1 and 2.



**Figure 1:** terms identified for those web sites that have *No* to *B8A*



**Figure 2:** terms identified for those web sites that have Yes to B8A

For what concerns the results of our classification methods, it is possible to consider the confusion matrix in table 2 for the CA applied to the remaining 3041 web sites.

**Table 2:** Confusion matrix for the web sites (test set) classified with Correspondence Analysis.

| <b>Original classification</b> | <b>Predicted</b> |            | <b>Total</b> |
|--------------------------------|------------------|------------|--------------|
|                                | <b>No</b>        | <b>Yes</b> |              |
| No                             | 2418             | 357        | 2775         |
| Yes                            | 131              | 135        | 266          |
| <b>Total</b>                   | <b>2549</b>      | <b>492</b> | <b>3041</b>  |

The result of this classification method shows a success rate equals to 83.9%, while the sensitivity is 50.7% and the specificity is 87.1%. For all the other methods that were used, it is possible to consider directly these last indexes as in table 3.

**Table 3:** Success rate, sensitivity and specificity for all the classification methods used on the test set

| <b>Method</b>          | <b>Success rate</b> | <b>Sensitivity</b> | <b>Specificity</b> |
|------------------------|---------------------|--------------------|--------------------|
| Support vector machine | 77.3%               | 1.6%               | 99.1%              |
| Classification tree    | 78.3%               | 7.2%               | 98.7%              |
| Boosting               | 77.7%               | 40.3%              | 88.5%              |
| Random forest          | 81.3%               | 32.2%              | 95.3%              |
| Bagging                | 80.6%               | 34.7%              | 93.9%              |
| Consensus              | 81.7%               | 27.8%              | 97.2%              |
| Naïve Bayes            | 73.9%               | 50.2%              | 80.6%              |
| Naïve Bayes (reduced)  | 79%                 | 51.3%              | 86.5%              |

It has to be noticed that, with the term *Naïve Bayes (reduced)*, we refer to a method in which we further selected only those terms that appears to be highly associated to the answer *Yes* of *B8A*. To determine such association we used the *Chi-square* by considering only the 5% of the terms with a greater value of it.

## 5 Conclusions and future work

In this study we investigated the possibility of getting and managing a huge amount of data available through the web sites of enterprises; this represents an enrichment of the traditional information collected by ISTAT and it could offer a method to confirm or validate it in some cases.

We observed that the results obtained in the traditional way (questionnaires) and by the automatic approach could be different due to the peculiarities of the question and how it was interpreted by the respondents or by the automatic tools we implemented. In particular, the enterprises answering ‘yes’ to the web sales could refer to a web site of their property or to something managed by others.

Although this represents a possible limit of this experimental study, we consider satisfying the resulting classification (as shown by the success rate of 83.9% in table 2).

In order to increase the accuracy of the experimented approach, it will be necessary to investigate the *false positives* (cases in which the methodology finds an web sale that was not declared by the enterprise). These events could come from an inadequate comprehension of the question or from an inaccurate correspondence between the terms identified. They could, also, derive from the different period in which this study and the original survey were made

For us, this is a first step towards the adoption of Internet as a data Source for the production of Official Statistics; the obtained results are very promising with respect to their implementation in short/medium term, for what concerns the replication of the methodology in other contexts or for a constant measurement of the phenomenon. Furthermore, it could be applied on the whole universe of enterprises (and not just in a sample of these), and it could permit the identification of a subpopulation of those active in web sales that can constitutes a new sampling frame, or even as a support to the survey for data correction operations.

## References

1. HLG-BAS (2011). Strategic vision of the High-level group for strategic developments in business architecture in statistics. Conference of European Statisticians, 59th Plenary Session. URL: <http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2011/1.e.pdf>
2. UNECE (2013). What does Big data mean for official statistics? Working paper. URL: <http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170622>
3. Scheveningen Memorandum (2013): [http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp\\_ess/0\\_DOCS/estat/SCHEVENINGEN\\_ME MORANDUM%20Final%20version\\_0.pdf](http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SCHEVENINGEN_ME MORANDUM%20Final%20version_0.pdf)
4. DIALOGIC (2010): Internet as data source Feasibility Study on Statistical Methods on Internet as a Source of Data Gathering, EU Commission Study. URL: [http://ec.europa.eu/information\\_society/newsroom/cf/itemdetail.cfm?item\\_id=8701](http://ec.europa.eu/information_society/newsroom/cf/itemdetail.cfm?item_id=8701)
5. CBS (2010): Rutger Hoekstra, Olav ten Bosch and Frank Harteveld . Automated Data Collection from Web Sources for Official Statistics: First Experiences. The Hague: CBS
6. Jsoup is an open source java project distributed under the liberal MIT license library for working with real-world HTML; it can be freely downloaded at: <http://jsoup.org/>
7. ADaMSoft is a free and open-source system, written in Java, for data management and data mining; it can be freely downloaded from: <http://adamsoft.sourceforge.net/>
8. TreeTagger is a tool for annotating text with part-of-speech and lemma information; it can be freely downloaded from: <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>
9. Text Mining with ADaMSoft: <http://sourceforge.net/projects/adamsoft/files/Guides/ADaMSoftTextMining.pdf/download>