

# Internet as a Data Source

## Scenari di uso di dati raccolti da Internet per la produzione statistica

Monica Scannapieco(\*)

28/5/2015

Lavoro condotto con:

**Giulio Barcaroli(\*) Alessandra Nurra(\*)**

**Sergio Salamone(\*)**

**Marco Scarnò(\*\*) Donato Summa(\*)**

(\*) Istituto Nazionale di Statistica (Istat)

(\*\*) Cineca



# Outline

1. What's the meaning of Internet as a Data Source?
2. Internet Data: How to access them?
3. Data Extraction pipeline
4. Internet as a Data Source for Official Statistics: Examples
5. «ICT Usage by Enterprises and Public Administration» Survey
6. Conclusions

# What's the meaning of Internet as a Data Source?

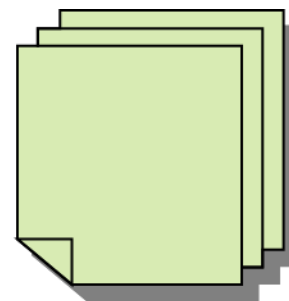
## Internet as a Data Source (IaD)

- Internet data can be used to (i) complement or (ii) substitute traditional data sources



AND/OR

Survey data



Administrative data



# Internet Data: How to access them?

Based on the actual location of the measurement, three basic types of IaD-methods can be distinguished:

- **User centric** measurements that capture changes in behaviour at the client (PC, smartphone) of an individual user. E.g. Browser visits monitors
- **Network-centric** measurements that focus on measuring properties of the underlying network. E.g. Traffic type monitors
- **Site-centric** measurements that obtain data from Web servers. E.g. Web sites scrapers



*European Commission: Internet as data Source - Feasibility Study on Statistical Methods on Internet as a Source of Data Gathering, 2012.*

# What is Web Scraping?

A step of the Data Extraction Pipeline

**Crawling**



**Scraping**



**Indexing**



**Searching**



# Data Extraction Pipeline: Crawling

## Crawling



- Crawling: a Web crawler (also called Web spider or ant or robot) is a software program that systematically browses the Web starting from (i) an Internet address (or a set of Internet addresses) and (ii) some pre-defined conditions, namely:
  - how many links navigate in depth
  - types of files to ignore
  - ...

# Data Extraction Pipeline: Scraping

## Scraping



- Scraping: a scraper takes Web resources (documents, images, etc.), and engages a process for extracting data from those resources, finalized to data storage for subsequent elaboration purposes.

# Data Extraction Pipeline: Indexing

## Indexing

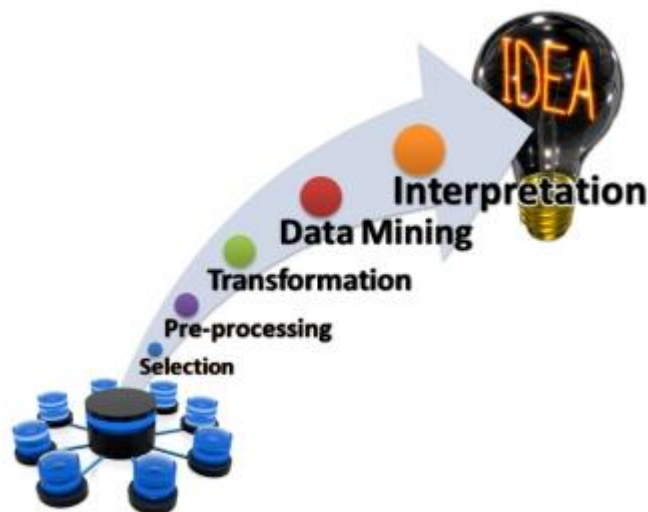


- Indexing: searching operations on a huge amount of data can be very slow, so it is necessary to index contents
  - Preliminary text processing operations to enable indexing



# Data Extraction Pipeline: Searching


## Searching



- Searching: Getting value from data through analysis methods

# Internet as a Data Source for Official Statistics: Examples

- Consumer Price Index (CBS- Netherlands)
  - Air Tickets
  - Property market
  - Clothes
- Consumer Price Index (Istat)
  - Consumer electronics
  - Air Tickets
- ICT Usage in Enterprises and Public Administrations (Istat)



Test if possible to make no  
assumption on the structures of  
Web sites

# The ICT Usage Project

Project supervised by the Istat Big Data Commission from February 2013 to February 2015

- **Purpose:**
  - Evaluate the possibility of adopting Web scraping and text mining techniques for estimates on the usage of ICT by enterprises and public institutions
- **Actors involved in the project:**
  - Istat, Survey on ICT usage in Enterprises and Public Institutions
  - Cineca
- **Outcomes:**
  - Starting from a supervised task, learning a model to be applied for estimating some answers to current questionnaires directly from Web sites

# The ICT Usage in Enterprises and Public Institutions (in short ICT Usage)

**B7** Indicare se l'impresa ha un proprio sito Web o home page, ovvero una o più pagine su Internet:

Sì ☐ 1

No ☐ 2

**B8** Indicare quali servizi offre il sito Web dell'impresa (*rispondere ad ogni riga*):

	Sì	No
a. Possibilità di effettuare ordinazioni o prenotazioni on line (es. carrello della spesa on line),....	1 <input type="checkbox"/>	2 <input type="checkbox"/>
b. Tracciabilità on line dell'ordine.....	1 <input type="checkbox"/>	2 <input type="checkbox"/>
c. Accesso a cataloghi di prodotti o listini prezzi.....	1 <input type="checkbox"/>	2 <input type="checkbox"/>
d. Possibilità di personalizzare i contenuti del sito per i visitatori abituali.....	1 <input type="checkbox"/>	2 <input type="checkbox"/>
e. Possibilità per i visitatori del sito di personalizzare o progettare prodotti.....	1 <input type="checkbox"/>	2 <input type="checkbox"/>
f. Avvertenze sulla politica in materia di privacy, marchio di certificazione della tutela della privacy o certificazione della sicurezza del sito.....	1 <input type="checkbox"/>	2 <input type="checkbox"/>
g. Annuncio di posti di lavoro vacanti o possibilità di effettuare domande di impiego on line,.....	1 <input type="checkbox"/>	2 <input type="checkbox"/>

## The “ICT Usage” survey

- In Italy, the survey investigates on a universe of 211,851 enterprises with at least 10 employees, by means of a sampling survey involving 19,186 of them (2011)
- In the 2013 round of the survey, 8,687 indicated their website (45% of sampling respondent units)
- The access to the indicated websites in order to gather information directly within them, gives different opportunities

# Phase 1: Crawling + Scraping

- Three systems tested:
  - The Apache Stack: Nutch/Solr, Nutch is a highly extensible and scalable open source web crawler. Solr is an open source enterprise search platform that is built on top of Apache Lucene
  - HTTrack, HTTrack is a free and open source software tool that permits to “mirror” locally a web site, by downloading each page that composes its structure.
  - JSOUP, open source Java library for working with real-world HTML.

# Phase 1: Crawling + Scraping

- Efficiency

Tool	# websites reached	Average number of webpages per site	Time spent	Type of Storage	Storage dimensions
Nutch	7020 / 8550=82,1%	15,2	32,5 hours	Binary files on HDFS	2,3 GB (data) 5,6 GB (index)
HTTrack	7710 / 8550=90,2%	43,5	6,7 days	HTML files on file system	16, 1 GB
JSOUP	7835/8550=91,6%	68	11 hours	HTML ADaMSoft compressed binary files	500MB

# Phase 1: Crawling + Scraping

- Effectiveness

<b>Tool</b>	<b>Access to specific element of HTML Pages</b>	<b>Download site content as whole for semantic extraction and discovery</b>	<b>Document Querying</b>	<b>Scalability to Big Data Size</b>
Nutch	Difficult	Easy	Easy	Easy
HTTrack	Easy	Easy	Difficult	Difficult
JSOUP	Easy	Easy	Difficult	Difficult

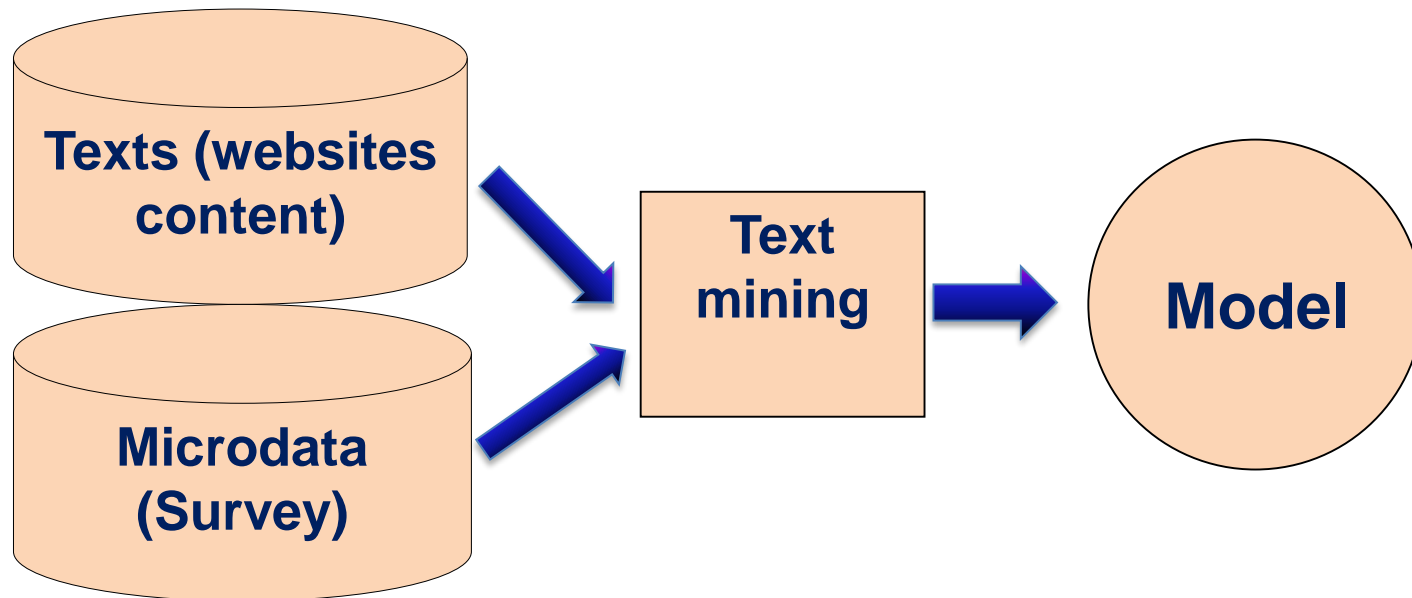


## Phase 1: Main Findings & Current Work

- Main Findings:
  - First example of scraping in OS without any assumption on the structure of the websites
  - Ability to scale up to a huge number of them
- Current Work:
  - Deployment and execution of Adamsoft/JSOUP and Nutch on CINECA PICO platform (1,080 cores, 54 nodes, 6.9 TB RAM)
    - <http://www.cineca.it/en/news/pico-cineca-new-platform-data-analytics-applications>

## Phase 2: Text Mining

- Predictive approach: subset of data related to sampled respondent units can be considered as the labeled data, and supervised learning methods can be applied
- In other words, the subset of 8,687 enterprises that indicated to have a website or a home page can be considered as the *training and test set* by means of which different models can be estimated in order to predict answers to [B8a : B8g] questions for the whole reference population.



# Evaluation of predictive models

*Application of different learners to predict question B8a “Online ordering or reservation or booking (Yes/No)”*

Learners	Indicators				
	Precision	Sensitivity	Specificity	Proportion of web sales functionality (observed)	Proportion of web sales functionality (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 aggregating	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
Naive Bayes	0.78	0.50	0.86	0.21	0.21

Accuracy=  
TP+TN/Total

TP / (TP+FN)

TN / (FP + TN)

## Phase 2: Main Findings and Current Work

- Also applied: Content analysis [Hopkins and Kings 2010]
- Main findings
  - Results are promising but still work to do to refine them
- Future work
  - predictors will be applied to the websites owned by all the enterprises in population of reference (about 100,000), in order to predict values at individual level and produce estimates by aggregating them
  - Evaluate the Mean Square Errors in the two cases (sampling estimates and estimates obtained in this full predictive approach) and to compare them: the reduction gained in terms of sampling errors will be less than or greater than the amount of bias due to the models applied?

# Conclusions

laD is a promising source for OS:

- Some scenarios are more easy to be put in production, e.g. ad-hoc scraping for online prices
- Others are more challenging, e.g. ICT Usage by enterprises, though promising anyway

