doi.org/10.26398/IJAS.0031-006

ONLINE JOB VACANCIES IN THE ITALIAN LABOUR MARKET

Ilaria Vannini[1]

ISTAT, Rome, Italy

Daniela Rotolone[2]

ANPAL Servizi, Rome, Italy

Cristian Di Stefano[3]

ISPRA, Rome, Italy

Abstract*. The analysis of online job vacancies is a research area that may provide a better understanding of the labour demand trends. With the aim of setting up an Italian vacancy monitor, about 70,000 job posts were collected for a three-month timeframe. By implementing the typical phases of a big data process, this work describes a procedure for the definition of a model able to integrate, represent and analyse the acquired data, focusing on the text-mining techniques of information retrieval and information extraction. In presenting the main results of the analysis, particular focus has been reserved to the description of the most challenging features as well as the operative solutions that characterised this activity, in order to ease a future extension of this methodology to a broader web job market.*

Keywords*: job vacancies, text-mining, big data, vacancy monitor.*

1. INTRODUCTION

Nowadays, online job advertisement is one of most popular channels to let labour supply meet labour demand. A job advertisement posted online can be replicated on several websites and different platforms, through employment-related search engines for job listings. Indeed, on the web, there are millions of job posts that could be used to analyse the Italian labour market and, specifically, labour demand trends. This could be done through the collection of the whole set of texts – *corpus* – belonging to the online job advertisements. Job posts are easy to identify, however their acquisition and analysis is challenging. As a matter of fact, they are characterised

[1] ilavannini@gmail.com
[2] drotolone@anpalservizi.it
[3] cristian.distefano@isprambiente.it

by a high variability of contents, not only because they can be collected from different platforms but also because, on the same platforms, different companies may post them. This requires dealing with many different technological formats and results in collecting different kinds of information. Moreover, they are not structured data, being written in «natural language text» and need to be transformed into structured, normalised database form.

From six job websites, about 70,000 employment ads have been collected between July and September 2015, and stored in a structured database, with the main goal of testing a process of data acquisition, extraction and analysis for the future set up of a job vacancy monitor, able to count them and monitor the job market in a more cost effective and timely manner than surveys based on other mass-media or sample surveys.

This paper aims at describing the whole big data process, from the phases of acquisition, information extraction, cleaning and integration of web job advertisements to the one of data analysis.

Section 2 provides a theoretical framework for online job advertisements. Section 3 illustrates the methodological framework and text-mining tools – information retrieval and information extraction –. Section 4 and Section 5 show, respectively, the web acquisition activities and the phase of data extraction and integration, while Section 6 is focused on data analysis and is followed by some concluding remarks.

2. ON-LINE ADVERTISEMENTS

A lot of job offers can be found across the web, which mainly differ from each other depending on the nature of the online recruitment agencies publishing them: they span from renowned and structured companies, such as Indeed and Monster, or small firms willing to recruit individuals by leveraging free job-posting sites, such as Bakeca.it, Subito.it, Kijiji.it etc.

Employment ads are the best way companies have to let everybody know what they need and to make job seekers aware of the appropriateness of their application (Jones *et al.*, 2006; Maurer *et al.*, 2007; Walker *et al.*, 2013). They are typical examples of company communication and can be seen as a part of an «embedded social process that over time produces, reproduces, and modifies particular genres of communication» (Yates *et al.*, 1992); they can be considered as «organisational artefacts that enact or celebrate organisational processes (…). Employment ads may therefore be a medium that connects individuals, groups, occupations, and organisations» (Rafaeli *et al.*, 1998, p.343). A job advertisement may be analysed

from many different point of views (Rafaeli *et al*., 1998): (1) the individual one, because it is a vehicle of attraction and recruitment of individuals; (2) the occupational one, i.e. the wording used is meaningful only in the context of the particular job advertised; (3) the organisational one, indeed its rhetoric contains explicit references to the organisation involved in the recruitment process; (4) the social one, making explicit reference to the company in which the job is offered and (5) the one of the specific field of occupation, i.e. it explicitly classifies the organisation involved into industry.

As a matter of fact, recruitment could be used for marketing campaigns, such as the case of electronic industry, where job descriptions focus mainly on the development of products, service innovations and technological growth and little information is given about benefits, career opportunities, cultural or working environment. Sales and service sectors, which are characterised by a high turnover, are more focused on workers, instead, and give potential applicants more information about aspects related with career and working environment.

Employment ads are often written using the specific language of the field of application, while sometimes they are written in a very informal way, lacking the main information a candidate would need to know for applying, such as the name of the posted job. It means that it is not obvious that job posts are clear, complete and informative (Bearden *et al*., 2006).

In general, job offers include information about four main elements: (1) the job name; (2) the company description; (3) the job description; (4) features that allow candidates to apply for the job.

Job description and job name are the most important fields among those listed above. The first contains information about the job location, the main characteristics of the company (for example economic sector), the type of contract offered, the skills asked to the applicant and the requisites needed to apply as well. The second one is the key element for counting and classifying the job vacancies available across the web.

Nowadays, thanks to technologies for the extraction of data from the web, it is possible to list all the existing vacancies, taking the process of their development into account, revealing which types of jobs are new and which are obsolete.

3. METHODOLOGICAL FRAMEWORK

The collection and analysis of job advertisements from the web is a challenging activity, which belongs to the field of «big data» analysis. As a matter of fact, millions of employment ads are available across the web. To analyse big data,

multiple distinct phases are required (Figure 1). Even if each of these phases is of the utmost importance, the focus is often on the analysis/modelling phase, which however is of little use without giving the proper attention to the previous activities of the pipeline. This is the reason why the present paper describes in depth the phases of acquisition, information extraction and cleaning and data integration, aggregation, and representation (Woods, 2011).

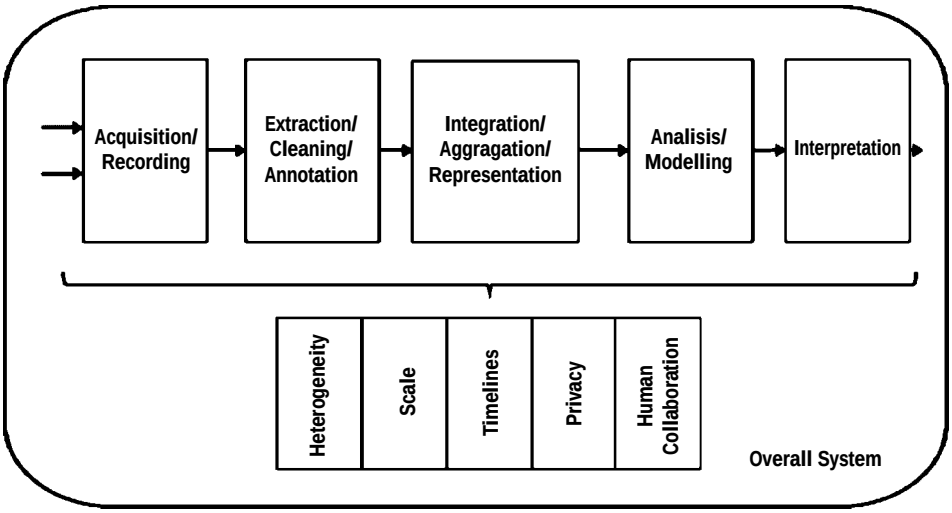

Fig. 1: The big data analysis pipeline (Jagadish, 2012)

The data acquisition phase is for acquiring information from a rich and varied data environment, the main objective being to filter irrelevant information and not discard useful one, automatically generating the metadata to describe what data is recorded and how. Information extraction and cleaning is needed to turn the information collected into a format suitable for analysis, The following phase of data integration, aggregation and representation, combines data acquired and extracted from several disparate sources and stored using various technologies, providing a unified view of them.

Being the unit of interest of the collection of job advertisement from the web made of textual documents, the development of the phases described above requires text-mining techniques. They are useful to extract contents from unstructured high dimension textual documents and to transform them in a structured database. Text-mining involves information retrieval (IR) and information extraction (IE) techniques. The first one refers to the finding of unstructured material, usually text documents, from within large collections to satisfy an information need and the second one to

the automatic extraction of structured information from unstructured one.

A text-mining project, generally, is made of three phases: (1) the gathering of unstructured information; (2) its transformation in structured or semi-structured information and (3) the application of analysis techniques and data mining for summarising the collected data into useful information.

An IR process is based on information needs formally written as queries, which result in several objects, such as text documents, images, videos and so on, depending on the application, with different degree of relevancy.

IE involves five major tasks: (1) segmentation, which allows to identify the text snippets that will fill a specific field of a database; (2) classification, which classifies the text segment in the right field of the database; (3) association, which identifies the fields to be put together in the same record; (4) normalisation, which puts information in a standard format and (5) de-duplication, which deletes duplicate records (McCallum, 2005).

Based on the complexity of the text sources, these tasks may be addressed in different ways: when there is sufficient formatting regularity, regular expression or hand-tuned, programmed rules are used, otherwise IE must rely on the language itself, such as the words, word order, grammar rules and on available irregular formatting clues. Moreover, at present, statistical and machine learning methods for IE have been used, allowing to automatically identify the rules for performing the tasks described above, showing the machine what to do on specific example texts and letting the machine generalises from these examples, tuning its own rules.

As for the collection of employment ads across the web, being in an experimental phase, this work was focused on the Italian labour market and six specific employment-related search engines for job listings. As a matter of fact, job vacancies are available on several job websites, social networks, blogs etc. Among these, for the acquisition of the job advertisements, the Italian agencies of some multinational companies – Monster, Adecco, Randstad, Men at Work (Maw), Indeed and Infojobs – have been selected.

Monster has been chosen because it is recognised as the main job search engine used in Italy. The others are notoriously the most relevant labour market agencies based in Italy. Such agencies, in addition to provide their intermediation services for the private sector, may also be authorised to do the same for public institutions, such districts and the Ministry of Labour, substantially providing the same services as the public employment centres provide. Besides this, such search engines, at the time this study was conducted, allowed a free extraction of their ads and proved to be consistent and stable over time, making them suitable to serve as sources for the following analysis.

4. ACQUISITION FROM THE WEB

The process of data acquisition from the web started on 15th July, 2015 and ended on 15th September, 2015. It was implemented with Python, a high level object-oriented programming language, with several software libraries among which BeautifulSoup and Requests were chosen. The first one is a library for pulling data out of HTML and XML files, while the latter is a simple HTTP library for Python, Apache2 licensed.

BeautifulSoup allows to convert the texts in an HTML document and gives them a structure, using the command: soup = BeautifulSoup (data, 'lxml'), where «data» can be an HTML file. Once the data structure «soup» is created, it is possible to ask to «Find all the links», or «Find the table heading that's got bold text, then give me that text».

In short, the acquisition phase was implemented through the following steps: (1) identification of data structures and instances of interest from a web site built with HTML and CSS and also written using JavaScript; (2) their extraction using the specific Python software/algorithm described in Section 4.1; (3) transformation and storing of the data in a proper structure for the purposes of querying and analysis and (4) loading transformed data into a PostgreSQL database, which is an open source object-relational database system (Figure 2).

Fig. 2: The acquisition phase process

4.1 HARVESTING AND CRAWLING

The main phases of data acquisition consisted of harvesting and crawling. During the harvesting phase, the developed software took the URL of the websites selected for this work (i.e. http://cercalavoro.monster.it/lavoro/ for Monster), scanned the content of the search results and extracted all the URLs directing to the detailed job offers (Figures 3 and 4).

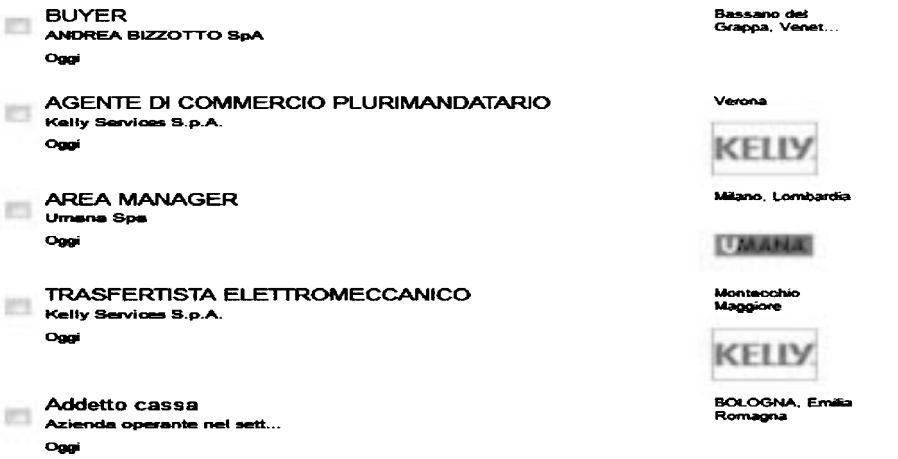

Fig. 3: Search results coming from Monster (step 1 of the acquisition phase process).

Fig. 4: Information extraction from HTML page through Beautiful Soup

Once all the URLs were collected and registered in a text file, the crawling phase started. More specifically, the crawler opened each link, extracted the required information and stored it in the database. The code of the crawler executed a series of iterative tasks, as shown below:

1 Scanning of the above mentioned text file and retrieval of an URL;
2 Check on the URL to avoid duplications;
3 Crawling of the detailed job offer;
4 Extraction of the relevant information: for each URL the software created a Python dictionary specular to the database record and filled it using the same methodology implemented for the harvesting code, that is locating and extracting interesting HTML tags (see example in Figure 5);
5 Storage in the database of data collected.

```
description = soup.find('div', {'itemprop':'description'}).text

record['description'] = description.replace('\n',' ')
```

Fig. 5: Example of extraction of «description» from job offer page and mapping into a dictionary field

At the end of the acquisition phase, the database contained different tables, one for each job site, with different fields depending on the information available in them. The following list covers all available fields: URL, job name, job description, type of contract, employment sector, job location (region, province and municipality), source of data, language of the offer, salary, years of experience, level of education, name of the company, number of positions offered, data of publication of the post. Database population was done through a SQL INSERT into the PostgreSQL table using the Python psycopg2 library (see example in Figure 6):

```
…..query = «INSERT INTO monster (url, titolo, descrizione, name, joblocation,
industry, employmentType, education, qualifications, dateposted, timestamp) VALUES
(%s, %s, %s, %s, %s, %s, %s, %s, %s, %s, %s);»data = (record['url'], record['titolo'],
record['descrizione'], record['name'], record['joblocation'], record['industry'],
record['employmentType'], record['education'], record['qualifications'],
record['dateposted'], time.ctime())cur.execute(query, data)

….
```

Fig. 6: Example of database population through a SQL INSERT into the PostreSQL using Python

5. DATA EXTRACTION AND INTEGRATION

After creating the database by applying the algorithms described in the previous section, the phase of data extraction and integration followed. It was implemented by an iterative process that included sources analysis (i.e. identifying what information was common to all advertisements across web sites and their degree of comparability), standardisation, cleaning, normalisation and review activities, aimed at enhancing the quality of the collected information as well as integrating missing information.

5.1 EXTRACTION, CLEANING AND ANNOTATION

The structure variety of employment ads emerged since the very first analysis of the sources, both across different job search platforms and within the same, because oftentimes users are free to enter or not information and sometimes they prefer to enter all information as free-text.

Job ads turned out to share very few common features: job name, job description, job location, kind of contract and economic sector or job related tasks. Although some of the analysed job ads included additional information, such as the required level of education or the base salary, it was not possible to add these fields to the database, given the poor quality (lack of completeness) of these data.

The phase of cleaning and annotation was the most complex and time-consuming. It included the identification and removal of:

(1) Job ads written in languages other than Italian, to facilitate the text-mining activities. To this regards, it can be noticed that the ads published in a foreign language on Italian web sites accounted for less than 1% of the total, mainly English and languages spoken by neighbouring countries, such as German and Slavic.

(2) Job ads lacking the most important information, such as the job description – « empty ads » – (around 2% of the total).

(3) Duplicated job ads, around 2% of the total, posted multiple times on the same site (for example, an expired ad that was republished as nobody had applied or no suitable candidates had been found; in such cases the ad was republished or its expiration date was postponed).

(4) Multiple job ads posting. The same ad was posted on several sources. In particular, quite often Monster aggregates ads published by others and, therefore, replicated ads were identified and discarded, for a percentage lower than 3% of the total.

(5) Non pertinent job ads, i.e. ads that are not actual job offers. By running a thorough text analysis of the job names, the types of contract and the job

descriptions, it emerged that 40% of the total number of all ads falls into this category: internships, training opportunities, short engagements (less than 3 days of work, such as staff searches for specific events like congresses) and hiring announcements for the public sector.

According to the Italian labour legislation, internships, that represent the vast majority of non pertinent jobs ads, are not recognised as regular work contracts and they are not considered in the employment calculation. For such a reason internships were not considered in this study.

Tab. 1: Number of ads, by job site, before and after the extraction and integration phase

Job site	After acquisition	After extraction and integration
Adecco	7,240	3,579
Indeed	17,690	8,719
Infojobs	21,856	10,911
Maw	2,344	2,341
Monster	13,048	4,740
Randstand	5,992	2,989
Total	68,170	33,279

Typing errors had been corrected and punctuations (i.e. commas, inverted commas, hyphen etc.) had been removed. Besides this, being each field populated with words written in different ways but with the same meaning, a phase of standardisation had been required, where these terms had been assigned unique values by applying manually-constructed synonym dictionary, simple rules and regular expressions (for example, for the type of contract, «part time» job was almost always used, but sometimes it could appear as «morning only» job).

A consistency check had been performed on every field of the database, to ensure that, when filled in, the field contained the proper information (avoiding that, for example, instead of the information about the type of contract, the related field would contain information about the job schedule).

Finally, inferring the information from the free text job description, missing values fields had been populated.

5.2 INTEGRATION AND AGGREGATION

The integration and aggregation phase, which involved the type of contract, the job location and the economic sector, had been developed from bottom-up activities and in the following phases: (1) check of the availability of the information among the different sources; (2) identification of the dictionary used by different sources for the same field; (3) mapping of the dictionaries to identify a common one

by clustering activities; (4) recodification of texts, to collapse each relevant word into the respective term of the new glossary and (5) enhancement of data coverage through deterministic imputation of missing or incomplete data.

As for activity (3), because of the diversity of dictionaries used by each source of data, an official classification had been chosen, i.e., for the type of contract, the official classification of the Ministry of Labour and Social Affairs[4] and, for the job location, the official classification of Italian municipalities, provinces and regions. Moreover, with regards to the place of work, the available information had proven to be oftentimes partial and did not allow a proper identification of the place. For example, the word «Roma» occurred in the name of 44 towns across Italy or sometimes the ads were integrated with additional non-coded information, such as area or neighbourhood, in an effort to better identify the place, especially when large towns are involved. This information had been analysed and used to correctly identify the place of work.

Data regarding the industry sector of employment, whose dictionary strictly depends on the professional specialisation of the job web sites, had been clustered to find a common glossary.

All of the activities described above contributed to guarantee an overall consistency to all the fields of the database, regardless of the source platform from which they had been collected.

The coverage level achieved at the end of this phase increased significantly if compared to the initial one, reaching almost the 100% for every field. In most of the cases, the information was retrieved from the free text job description.

Tab. 2: Percentage coverage of the fields related the type of contract, job location and sector of employment, before and after extraction and integration phase

Source of data	Type of Contract		Job location		Sector of employment	
	Before	*After*	*Before*	*After*	*Before*	*After*
Adecco	90	100	90	100	100	100
Indeed	26	99.98	99	99	0	0
Infojobs	97	51	97	100	0	0
Maw	98	100	98	100	77	100
Monster	4	100	37	100	53	100
Randstand	84	99.99	84	100	99	100

[4] With regards to the type of contract, it should be pointed out that most of the times each source implements its own dictionary of terms, too. Some use terms like: «Type of job: temporary», some just «Temporary», some allow users to specify this information in a free-text specific field.

6. DATA ANALYSIS

In general, the analysis of job posts by text-mining techniques aims at the identification of the core skills required for a position (Bolasco et al, 2004) and to detect similarities among ads through clustering, regardless of their job name. (Aureli *et al*., 2006; Iezzi *et al*., 2013; Iezzi *et al*., 2011). The main objective of the analysis of the data collected above, a total of 33,279 ads classified by source, job name, job description, type of contract, job location and sector of employment, was to distinguish job advertisements' main characteristics and to find their «lexical worlds» (Bolasco *et al*., 2004).

6.1 DESCRIPTIVE STATISTICS

Most of the ads was extracted from Infojobs (32,8%) and Indeed (26,2%), followed by Monster (14,2%), Adecco (10,8%), Randstad (9%) and Maw (7%) (Table 3).

Tab. 3: Distribution of ads by source

Source	Frequency	%
Adecco	3,579	10. 8
Indeed	8,719	26.2
Infojobs	10,911	32.8
Maw	2,341	7.0
Monster	4,740	14.2
Randstand	2,989	9.0
Total	**33,279**	**100**

Jobs offered were mainly based in the north of Italy (40.8% north-west and 31.9% north-east). Ads for positions based in the centre of Italy were 17.3%, while ads for positions based in the south of Italy and islands were respectively 6.5% and 2.8%. Moreover, there were few ads generically based in the whole of Italy and abroad. (Table 4).

Tab. 4: Distribution of ads by geographical breakdown

Geographical breakdown	Frequency	%
Northwest Italy	13,575	40.8
Northeast Italy	10,514	31.6
Central Italy	5,763	17.3
Southern Italy	2,161	6.5
Insular Italy	920	2.8
Total	**32,933**	**99.0**
Italy	270	0.8
Abroad	76	0.2
Total	**33,279**	**100**

The distribution of ads by geographical breakdown is highly correlated with that of the total number of workers in the third quarter of 2015, the correlation coefficient being higher than 0.96, accordingly with the cross-sectional quarterly data coming from the Labour Force Survey collected by National Institute of Statistics (ISTAT) (Table 5).

Tab. 5: Comparison between the total number of workers and the number of ads by geographical breakdown

Geographical breakdown	Number of workers (3rd quarter 2015)	Number of ads (15th June – 15th September 2015)
Northern Italy	11,707,281	24,089
Central Italy	4,871,687	5,763
Southern Italy	5,973,529	3,081
Total	**22,552,496**	**32,933**

The 45.9% of jobs offered were mainly atypical (temporary contracts or contracts for the provision of independent services). Instead, in the case of contracts for dependent employment, most of them were fixed-term (26%) versus 11.2% of ads offering permanent contracts (Table 6).

Tab. 6: Distribution of ads by type of contract

Types of contract	Frequency	%
Fixed-term contract	8,657	26.0
Permanent contract	3,735	11.2
Atypical contract	15,277	45.9
Not specified	5,610	16.9
Total	33,279	100

According to data from the Labour Force Survey, the number of employed with permanent contracts was higher in the third quarter of 2015 than before, meaning that the labour market demand of job had been met (with an increase of about 73,000 employed with fixed-term contracts from the second to the third quarter of 2015).

As for the sectors of activity, the majority of ads referred to positions in the field of information technology and communication or for specialised workers (Table 7).

Tab. 7: Distribution of ads by sector of employment

Sector	**%**
Administrative, Accounting, Secretary	1.0
Automotive	0.5
Banking/Financial	1.2
Consumer care	0.5
Real estate	0.5
Energy	0.6
Large-scale retail channel, Retail	2.8
Chemical and pharmaceutical«industry	0.5
Information technology & telecommunication	3.4
Engineering	2.5
Mass media and show business	0.8
Specialised worker	9.7
Recruitment	0.5
Sales	0.7
Healthcare	0.8
Logistics and transportation industry, safety	1.7
Tourism and catering	0.9
Other (<0,4%)	2.9
Not classified	7.1
Total* (N. 12,708)	**38.2**

* *Missing data for Indeed e Infojobs*

6.2 TEXTUAL ANALYSIS

This section describes the textual data collected, which represent a *corpus* composed of 3,105,419 word tokens and 41,830 word types, with the objective of finding the main patterns for the classification of job ads.

The most frequent words appearing in the vocabulary, with the exception of stop-words, that is the commonly used words not corresponding to any particular subject matter, and of theme words for the *corpus*, mainly referred to the employment sector and required skills, such as sales, management, language (good/excellent English). (Table 8).

Tab. 8: The most frequents words of the vocabulary

Word	freq.	Word	freq.
esperienza (experience)	28,214	gestione (management)	10,919
azienda (company)	27,873	risorsa (resource)	10,842
cliente (consumer)	20,364	disponibilità (availability)	10,825
settore (sector)	18,806	offrire (to offer)	10,345
ricerca (search)	17,643	buono (good)	10,221
conoscenza (knowledge)	16,842	requisito (requirement)	10,140
candidato (candidate)	16,704	ottimo (excellent)	9,637
contratto (contract)	15,787	offerta (offer)	9,616
richiedere (ask for)	15,670	servizio (service)	9,375
vendita (sales)	11,834	filiale (branch)	9,352

Moreover, by extracting the meaningful parts of the vocabulary[5], namely the «peculiar language», which identifies words under/over-used with respect to the expected use according to a frequency dictionary of reference (in this case, the standard Italian[6]), keywords of the documents were detected. They refer to: legal aspects of recruitment and hiring; the sector of employment (i.e. telemarketing, selling, engineering); skills (i.e. relational, office, English); qualification (i.e. bachelor degree or high school diploma) and the type of contract. (Table 9).

Tab. 9: Keywords of the documents (peculiar language with respect to the standard Italian)

Word	freq.	Word	freq.
ricerchiamo (search)	5,627	esperienza (experience)	26,567
privacy	3,679	seleziona (to select)	3,301
mansione (task)	6,300	team	3,907
somministrazione (supply contract)	4,672	sito (website)	3,862
occuperà (will take responsibility for)	5,173	metalmeccanico (steelworker)	2,723
filiale (branch)	9,094	metalmeccanica (metal)	2,134
oraria (time slot)	3,222	trasferte (business trips)	1,675
aut (authorisation)	4,550	risorsa (resource)	8,121
requisiti (requirements)	9,197	maturato (work experience)	4,242
sessi (sexes)	4,727	relazionali (interpersonal skills)	2,425

5 The analysis of peculiar and specific language was carried out using Taltac.

6 A *corpus* set up from different sources (written and spoken language, amounting to four millions of occurrences), which is able to define the prevalent use of words in the standard Italian language, http://www.taltac.it/it/taltac210.shtml.

Then, the analysis of words under/over-used for each source of data, with respect to a value of reference (the average use) allowed to identify the sources «specific language». It was clearly related to the specificity of the different job search engines: Adecco focused on engineering, pharmaceutical and chemical industry; Indeed mostly published ads for receptionists, catering services and secretaries; Infojobs focused on marketing and sales; Maw and Randstad addressed technical, specialised and highly skilled positions and Monster the information technology sector. (Table 10).

Tab. 10 (part a): Specific language according to the source of data – first fifteen significant words in alphabetical order

Adecco	Indeed	Infojobs
auto (car)	badare (caregiver)	ambizioso (challenging)
chimico (chemical)	barista (barman)	commerciale (business)
dermocosmetico (skin-care)	cameriere (waiter)	crescita (growth)
elettrotecnico (electrical engineer)	cassa (cash desk)	formazione (training)
farmaceutico (pharmaceutical)	cassiere (cashier)	giornaliero (daily)
farmacia (pharmacy)	maturità (high school diploma)	Internet
full time	negozio (shop)	Marketing
industria (industry)	pulizia (cleaning)	meritocratico (meritocratic)
manifatturiero (manufacturing)	puntualità (punctuality)	provvigione (commission)
metalmeccanico (steelworker)	reception	serenità (composure)
Office	receptionist	tablet
patente (driving licence)	segretario (secretary)	telemarketing
Sales	segreteria (secretaryship)	tranquillità (calmness)
somministrazione (supply contract)	serietà (reliability)	vendita (sale)
Technology	vitto (room and board)	viaggio (travel)

To better understand the analysis of the phenomenon and to identify the main patterns of online job ads grouping them by similarity, a cluster analysis of job post descriptions was carried out, applying a descending hierarchical classification using the Alceste procedure (Reinert, 1998). The algorithm allows dividing the *corpus* into different classes, by decomposing the text into groups of words or sentences, i.e. the elementary contextual units (ECUs), and grouping them into lexical classes as a function of the co-occurrence of words composing these ECUs. Thus, most similar words are grouped in the same clusters and the most different words fall in different ones. The procedure is divisive: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Each class of words extracted using this method is related to a specific theme, which must be interpreted by the researcher. To make their interpretation easier,

region, sector, source and type of contract have been used as classification variables in the analysis.

Table 10 (part b): Specific language according to the source of data – first fifteen significant words in alphabetical order

Maw	Monster	Randstand
autonomia (authonomy)	analisi (analysis)	buono (good)
carpenteria (carpentry)	Business	disegno (drawing)
disegnatore (designer)	development	inglese (English)
disegno (drawing)	engineer	legno (wood)
IT	html	lingua (language)
legno (wood)	ICT	macchina (car)
macchina (car)	ingegneria (engineering)	meccanico (mechanic)
macchinario (machine)	Java	metalmeccanico (steelworker)
meccanico (mechanic)	manager	montaggio (assembling)
metalmeccanico (steelworker)	Oracle	operaio (worker)
perito (valuer, expert)	progettazione (planning)	promotore (promoter)
produzione (production)	software	prospettiva (outlook)
saldatura (weld)	Sql	saldatore (welder)
stampo (mould)	tecnologia (technology)	scopo (goal)
trasferta (business trip)	tecnologico (technological)	technical

We were able to build five clusters, corresponding to five «lexical worlds» (Reinert, 1998), concerning: (1) the kind of contract offered and the related characteristics and prerequisites (21.7% of job ads); (2) engineering technical and specialised skills (17.9% of job ads); (3) sales (20.3% of ads); (4) software technical and specialised skills (24.9%); (5) legal aspects (15.2%).

The most characteristic words for each cluster are represented in Figure 7, together with the classification variables' values.

The results of the analysis have provided five clusters offering interesting insights of the job offers sector. These clusters show the main patterns that characterise the job offers on the web and confirm the outcomes of the job offers surveys carried out through other channels, as well as sector studies related to the jobs market.

Cluster #1 confirms that the job offer is generally accompanied by the typical attributes of the job offer: the contract type (permanent/temporary, full-time/part-time) first of all, but also other characteristics that contribute to define all aspects of the job offer, such as the availability of the candidates to travel, oftentimes driving their own cars.

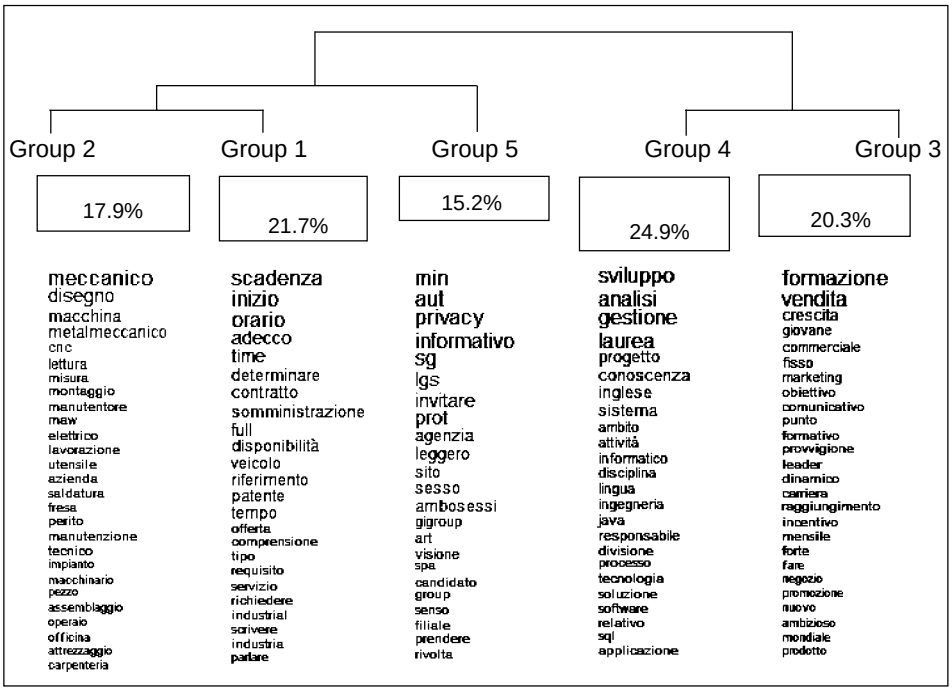

Fig. 7: The five groups identified by the cluster analysis

Cluster #2 shows that the most required professionals are the high-skilled mechanical/engineering workers.

Clusters #3 and #4 clearly show the most required professional skills in today's job market. Cluster #3 shows a leading pattern for job searches; this is the sales sector and the typical skills requested to the individuals working in this field: young, dynamic and result-oriented, with salaries that include a basic and a variable component, which depends on their performances.

Cluster #4 shows that software engineers are in high demand, especially developers (java/sql the most required coding languages) and professionals who can coordinate and support the whole life-cycle of a software product (we noticed words like analysis, development, management which represent the typical phases of the software life cycle). Besides this, a professional working level knowledge of English is frequently mandatory to work in this field.

Cluster #5 shows some legal aspects of recruitment and hiring, identified by a few typical and very heterogeneous terms, that would deserve a more thorough analysis.

Finally, by analysing the job names, it was possible to identify the professions in higher demand on the web (Table 11). After the phase of integration, the field «job name» counts 13,273 different professions, which decreased to 5,785 after performing further data cleaning, required because of the presence of information other than the specific job offered – i.e. number of positions, job location, kind of contract – and the possibility to write the same job in different ways, too. However, further efforts are needed for addressing jobs called in different ways but related to the same activity. Obviously, this top ten ranking is strongly affected by the seasonality of the period, i.e. the summer of 2015.

Tab. 11: The top ten of the professions most in demand on the web

Profession	Frequency	%	% of total ads
Accountant	442	8.2	1.3
Employee	1,678	31.3	5.0
Mechanical engineer	231	4.3	0.7
Mechanical maintenance technician	206	3.8	0.6
Mechanical fitter	218	4.1	0.7
Workman	1,226	22.8	3.7
Mechanical designer	249	4.6	0.7
Cleaner	306	5.7	0.9
Sales business	235	4.4	0.7
Salesperson	577	10.7	1.7
Total	**5,368**	**100.0**	**16.1**

7. CONCLUSIONS

Being online job advertising the most popular channel to let labour supply meet labour demand, the analysis of online job posts is a relevant research area for a better understanding of the labour demand trends. Indeed, the collection of the whole set of texts – *corpus* – belonging to millions of job ads which proliferate across the web could be used to analyse the labour market. While job posts are easy to identify, their acquisition and analysis is instead challenging, mainly because of their structure variety, both in single and across multiple platforms, when users are allowed to enter information as free-text.

This work was aimed at developing a process to define a model capable of integrating and representing data retrieved from heterogeneous sources.

In this context, a process of big data analysis was implemented, focusing on the phases of acquisition, information extraction and cleaning, and data integration.

Through this process, a vacancy monitor for the Italian labour market was finally built, with the ultimate goal of determining to what extent web ads were real job offers, and identify their main characteristics, in terms of both basic and soft skills required.

The outcome of this process provided a set of structured data suitable to constitute the input for quantitative and comparative analysis.

Finally, the usage of text-mining techniques of information retrieval and information extraction in the processing of the job descriptions was instrumental in the comprehension of the labour demand trends and the insights of the required skills, showing the main patterns that characterise the job offers on the web.

It was clear, in the end, that the implemented process was able to provide timely and accurate results through its iterative and streamlined steps and, at the same time, capable of improving cost-efficiency with respect to traditional survey-based studies performed on other channels. In addition, this process proved to be highly scalable and, as such, ideal to be extended to a broader job market, with no limitation to particular timeframe.

REFERENCES

Aureli, E. and Iezzi, D.F (2006). Recruitment via web and information technology: A model for ranking the competences in job market. *Proceedings of JADT 2006 (8es Journées Internationales d'Analyse Statistique des Données Textuelles)*. 1: 79-88.

Bearden, W. and Hardesty, D. (2006). Varying the content of job advertisements. The effects of message specificity. *Journal of Advertising*. 35 (1): 123–141.

Bolasco, S., Bisceglia, B. and Baiocchi, F. (2004). Estrazione di informazione dai testi. *Mondo Digitale*. 3(1): 27-43

Iezzi, D.F., Mastrangelo, M. and Sarlo, S. (2013). New fuzzy method to classify professional profiles from job announcements. In P. Giudici, S. Ingrassia and M. Vichi, *Statistical Models for Data Analysis*. Springer-Verlag, Berlin-Heidelberg: 151-159.

Iezzi, D.F., Mastrangelo, M. and Sarlo, S. (2011). Text clustering based on centrality measures: An application on job advertisements. *JADT 2011 (Statistical Analysis of Textual Data)*. 1: 515-524.

Jagadish, H.V. *Challenges and Opportunities with Big Data. A White Paper on Big Data*. http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf. Last access: 20/06/2015

Jones, D.A., Shultz, J.W. and Chapman, D.S. (2006). Recruiting through job ads. The effects of cognitive elaboration on decision making. *International Journal of Selection and Assessment*. 14 (2): 167-179.

Maurer, S.D. and Yuping, L. (2007). Developing effective e-recruiting websites. Insights for managers from marketers. *Business Horizons*. 50: 305–314.

McCallum, A. (2005). Information extraction: Distilling structured data from unstructured text. *Magazine Queue - Social Computing Queue Homepage Archive*. 3 (9): 48-57.

Rafaeli, A. and Oliver, A.L. (1998). Employment ads. A configurational research agenda. *Journal of Management Inquiry*. 7(4): 342-358.

Reinert, M. (1998). What is the object of a statistical analysis of discourse? Some reflections about the Alceste solution. *Proceedings of the 4th JADT (Journées d'Analyse des Données Textuelles).* JADT, Université de Nice JADT.

Yates, J. and Orlikowski, W.J. (1992). Genres of organizational communication. A structurational approach to studying communication and media. *The Academy of Management Review*. 17 (2): 299-326.

Walker, J.H. and Hinojosa, A.S. (2013). Recruitment. The role of job advertisements. In D.M. Cable and K.Y.T. Yu, editors, *Oxford Handbook of Recruitment*. Oxford University Press, New York: 269-283.

Woods, D. (2011). *Big Data Requires a Big Architecture*. Tech. Forbes.