Accepted Manuscript

Classifying online job advertisements through machine learning

Roberto Boselli, Mirko Cesarini, Fabio Mercorio, Mario Mezzanzanica

PII: S0167-739X(17)32183-0
DOI: https://doi.org/10.1016/j.future.2018.03.035
Reference: FUTURE 4047

To appear in: *Future Generation Computer Systems*

Received date : 27 September 2017
Revised date : 21 February 2018
Accepted date : 17 March 2018

Please cite this article as: R. Boselli, M. Cesarini, F. Mercorio, M. Mezzanzanica, Classifying online job advertisements through machine learning, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.03.035

Classifying Online Job Advertisements through Machine Learning

Roberto Boselli[a,b], Mirko Cesarini[a,b], Fabio Mercorio[a,b], Mario Mezzanzanica[a,b]

[a] *Department of Statistics and Quantitative Methods, University of Milan-Bicocca, Italy*
[b] *CRISP Research Centre, University of Milan-Bicocca, Italy*

Abstract

The rapid growth of Web usage for advertising job positions provides a great opportunity for real-time labour market monitoring. This is the aim of Labour Market Intelligence (LMI), a field that is becoming increasingly relevant to EU Labour Market policies design and evaluation. The analysis of Web job vacancies, indeed, represents a competitive advantage to labour market stakeholders with respect to classical survey-based analyses, as it allows for reducing the time-to-market of the analysis by moving towards a fact-based decision making model. In this paper, we present our approach for automatically classifying million Web job vacancies on a standard taxonomy of occupations. We show how this problem has been expressed in terms of text classification via machine learning. We also show how our approach has been applied to certain real-life projects and we discuss the benefits provided to end users.

Keywords: Machine Learning, Text Classification, Big Data, NLP

1. Introduction

In the last few years, the diffusion of web-centric services is growing exponentially, and this allows a significant part of the European Labour demand to be communicated through specialised web portals and services. This has also led to the introduction of the term "Labour Market Intelligence" (LMI), which refers to the use and design of AI algorithms and frameworks for Labour Market Data to support decision-making.

Motivating Example. In the on-line job market, a *job vacancy* is a document containing two main text fields: a *title* and a *full description*. The title shortly

summarizes the job position, while the full description field usually includes the position details and the relevant skills the employee must possess. Table 1 shows two job vacancies extracted from specialised web sites. Though both advertisements seek for *computer scientists*, the differences in terms of job requirements, skills and corresponding educational levels are quite evident. The first job vacancy (A) is looking for a software developer, while the second one (B) is looking for a less qualified candidate, i.e., an ICT technician. Indeed, in the latter case, the only requested abilities are to be able to *use* certain solutions, and the knowledge of a programming language (optional) which is usually taught in some professional high-schools.

Table 1: An example of Web job vacancies.

(A) **ICT Developer**. "Looking to recruit a software developer to join our dynamic R&D team to support work on Windows-based software for current and upcoming products. A flexible, self-starting attitude is essential along with a strong motivation to learn new skills as required. The ideal candidate will have at least two to three years experience working in a fast and dynamic small team. Experience in Python are key requirements for my client. A degree in computer science / computer engineering is preferable, but other engineering / science graduates will be considered if they have software development experience."	(B) **Application Consultant**. "We are seeking for an application consultant that will support our internal SW engineers in the realisation of ad-hoc ERP software. The ideal candidate should have (at least) a high level degree and 2+ years experience in using both Sharepoint-MS CRM and Magic. Coding skills on VB are appreciated."
Workplace: Milan	Workplace: Rome
Contract type: Permanent	Contract type: Unlimited Term

Being able to catch these differences and to classify these two distinct occupational profiles promptly and using a standard taxonomy is mandatory for analysing, sharing, and comparing the web Labour market dynamics over different regions and countries, focusing on the skills they require and linking them to the ones expected for such positions.

There is a growing interest in designing and implementing real LMI applications for Web Labour Market data that can support policy design and evaluation activities through evidence-based decision-making. In 2016, the European Commission highlighted the importance of Vocational and Educational activities, as they are "valued for fostering job-specific and transversal skills, facilitating the transition into employment and maintaining and updating the skills of the workforce according to sectorial, regional and local

needs"[1] In 2014, the Cedefop EU Agency[2] launched a call-for-tender[3] aimed at collecting Web job vacancies from five EU countries and identifying the requested skills from the data. The rationale behind the project is to turn data extracted from Web job vacancies into knowledge (and thus value) for policies design and evaluation through a fact-based decision making. In 2016, the EU launched the ESSnet Big Data project, involving 22 EU Member States with the aim of "integrating big data in the regular production of official statistics, through pilots exploring the potential of selected big data sources and building concrete applications".

Contribution. In this paper we describe how the problem of classifying and extracting useful knowledge from Web Labour Market Data has been addressed in terms of Text Classification and Information Extraction problems. To this end, we also show some results and outcomes in the field of Labour Market Intelligence (LMI) on two distinct research projects: *WollyBI*[4] and *Cedefop*[3], by focusing on (i) job vacancy classification and skill extraction tasks, and (ii) the support in decision making activities that our approach provided to the stakeholders involved. In summary, the contributions of this paper are the following:

- we build a machine learning model for classifying multilingual Web job vacancies, fully implemented into a system and a EU research project;
- we report experimental evaluation of machine learning algorithms employed in two real-life scenarios of the Web Labour Market;
- we show the significance of our approach on several research projects as well as the added value given to the labour market stakeholders involved, namely *Recruitment Agencies*, and *International Vocational and Educational Training Agencies*.

[1] The Commission Communication "A New Skills Agenda for Europe" COM(2016) 381/2, available at `https://goo.gl/Shw7bI`

[2] Cedefop European agency supports the development of European Vocational Education and Training (VET) policies and contributes to their implementation - `http://www.cedefop.europa.eu/`

[3] "Real-time Labour Market information on skill requirements: feasibility study and working prototype". Cedefop Reference number AO/RPA/VKVET-NSOFRO/Real-time LMI/010/14. Contract notice 2014/S 141-252026 of 15/07/2014 `https://goo.gl/qNjmrn`

[4] `www.wollybi.com`

Research's Goal. Figure 1 would summarise the main challenging issues we have identified while dealing with Web labour market information, and the stakeholders' needs that, in turn, have guided our research activities. The goal of this research line draws on the idea that reasoning over Web job vacancies represents an added value for both *public and private* labour market operators facilitating a deep understanding of Labour Market dynamics, occupations, skills, and trends by: (*i*) reducing the time-to-market cimpared to classical survey-based analyses (official Labour Market surveys results actually take up to one year to become available); (*ii*) overcoming the linguistic boundaries through the use of standard classification systems rather than proprietary ones; (*iii*) representing the resulting knowledge over several dimensions (e.g., territory, sectors, contracts, etc) at different levels of granularity and (*iv*) evaluating and comparing international labour markets to support fact-based decision making.

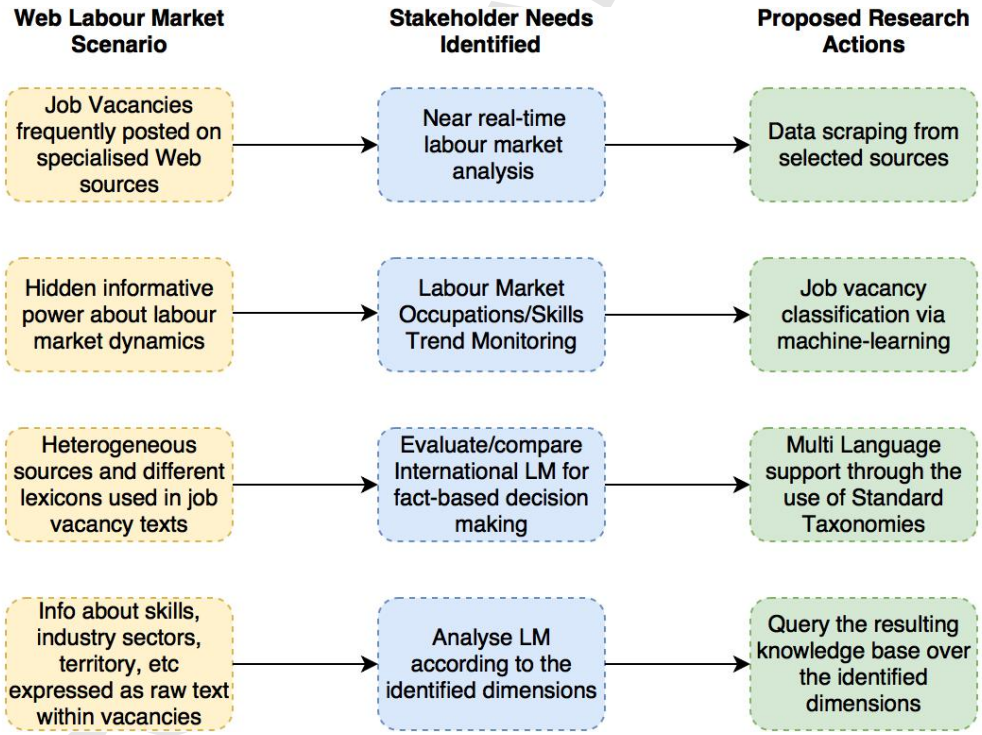

Figure 1: Schematic view of the main elements of (i) The Web LM scenario, (ii) some stakeholder's needs and (iii) the actions we propose to deal with.

2. Related Work

In this section we discuss the relevance of labour market analysis. As this is an *emerging* cross-disciplinary field of studies, here we distinguish among the *organisation*, *firm* and *literature* perspectives.

Organisation. There is a call for supporting the policy design and evaluation activities through evidence-based decision-making, as highlighted in the European Commission's Communication "New Skills for New Jobs",[5] and in one of the flagship initiatives of the Europe 2020 strategy, the "Agenda for new skills and jobs".[6] Furthermore, in 2010 the European Commission has published the communication "A new impetus for European Cooperation in Vocational Education and Training (VET) to support the Europe 2020 strategy",[7] aimed at promoting education systems in general, and VET in particular. In 2016 the European Commission has remarked the importance of VET's activities, as they are "valued for fostering job-specific and transversal skills, facilitating the transition into employment and maintaining and updating the skills of the workforce according to sectorial, regional and local needs".[8].

Firm. Focusing on the labour market domain, the extraction of meaningful information from unstructured texts has been mainly devoted to supporting the e-recruitment process (see, e.g.,[1]) attempting to support or automate the *resume* management by matching using machine learning approaches to match candidate profiles with job descriptions [2, 3, 4]. This problem is also relevant for business purposes, and it has motivated the development of several commercial products providing job seekers and companies with skill-matching tools. Concerning companies, their need to automatize Human Resource (HR) department activities is strong; as a consequence, a growing number of commercial skill-matching products have been developed in the last years, for instance BurningGlass, Workday, Pluralsight, EmployInsight, and TextKernel. To date, the only commercial solution that uses both ISCO and ESCO systems as thesauri is Janzz: a Web based platform for matching

[5]The Commission Communication "New Skills for New Jobs" (COM(2008) 868, 16.12.2008)

[6]The Commission Communication "An Agenda for new skills and jobs: A European contribution towards full employment" (COM(2010) 682, 23.11.2010)

[7]Publicly available at `https://goo.gl/Goluxo`

[8]The Commission Communication "A New Skills Agenda for Europe" COM(2016) 381/2, available at `https://goo.gl/Shw7bI`

labour demand and supply in both public and private sectors. It also provides APIs access to its knowledge base, but it is not aimed at classifying job vacancies. Worth of mentioning is Google Job Search API, a pay-as-you-go service announced in 2016 for classifying job vacancies through the Google Machine Learning service over O*NET, which is the US standard occupation taxonomy. Though this commercial service is still a closed alpha, it is quite promising and also sheds the light on the needs for reasoning with Web job vacancies using a common taxonomy as baseline.

Literature.. Since the early 90s, *text classification* (TC) has been an active research topic. It has been defined as "the activity of labelling natural language texts with thematic categories from a predefined set" [5]. Most popular techniques are based on the *machine learning* paradigm, according to which an automatic text classifier is created by using an inductive process able to learn, from a set of pre-classified documents, the characteristics of the categories of interest. The case in which one category must be assigned to each document is called *single-label* classification, while *multi-label* classification is the case when many categories may be assigned to the same document.

In the recent literature, text classification has proven to give good results in categorizing many real-life Web-based data such as, for instance, news and social media [6, 7], and sentiment analysis [8, 9]. To the best of our knowledge, text classifiers have not been applied yet to the classification of Web job vacancies published on several Web sites for analysing the Web job market of a geographical area, and the systemis the first example in this direction.

On the other side, skills extraction from Web job vacancies can be framed in the Information Extraction field [10, 11] and Named Entity Recognition [12]. The latter has been applied to solve numerous domain specific problems in the areas of Information Extraction and Normalization [13]. Worth of mentioning are also relevant works on skills extraction from resumes [13, 14, 15, 16].

The work described in this paper can be framed in the web mining field, according to [17] "Web mining aims to discover useful information or knowledge from Web hyperlinks, page contents, and usage logs". The interested reader can refer to [17, 18, 19] for more details on data mining and web data mining.

All these approaches are quite relevant and effective, and they also make evidence of the importance of the Web for labour market information.

Nonetheless, they differ from our approach in two aspects. First, we aim to classify *job vacancies* according to a target classification system for building a (language independent) knowledge base for analyses purposes, rather than matching resumes on job vacancies. Furthermore, resumes are usually accurately written by candidates whilst Web advertisements are written in a less accurate way, and this quality issue might have unpredictable effects on the information derived from them (see, e.g. [20, 21, 22, 23] just to cite a few). Second, our approach aims at producing analyses based on standard taxonomies (i.e., ISCO and ESCO) to support fact-based decision making activities of several stakeholders.

3. Preliminaries and Problem Formulation

A Web job offer can be seen as a document mainly composed of a pair of texts: a title and a (full job) description. The title summarises the working position offered by the employer, while the description usually provides the position details, including all the required relevant skills, according to the employer preferences. For our purposes we formalise this concept as follows.

Definition 1 (**web Job Vacancy**). *A Web job vacancy J is a 4-tuple $J = (i, s, t, d)$ where $i \in \mathbb{N}$ is a unique document vacancy identifier, s is an identifier of the Web source from which the job vacancy has been retrieved, t is the text describing the title while d is the full description of the job demanded.*

Unfortunately, the use of proprietary and language-dependent taxonomies can prevent the effective monitoring and evaluation of Labour Market dynamics across national borders. For these reasons, a great effort has been made by International organisations for designing *standard* classifications systems, that would act as a lingua-franca for sharing a machine-readable knowledge about the Labour Market and overcoming the linguistic boundaries as well. One of the most important classification systems designed for this purposes is ISCO: The *International Standard Classification of Occupations* has been developed by the International Labour Organization as a four-level classification that represents a standardised way to organise the labour market occupations. ISCO is a four-level hierarchically structured classification system that allows jobs to be classified into 436 unit groups. Formally speaking, it can seen as a tree where nodes are unit groups and directed edges between nodes are hierarchical "is-parent-of" relations. For

example, the job occupation *software developer* is classified over the ISCO pathway (2, "Professionals"), (2.5, "Information and Communications Technology Professionals"), (2.5.1, "Software and Applications Developers and Analysts"), (2.5.1.2, "Software Developers"). Clearly, the deeper the ISCO level, the higher the precision in classifying the job position.

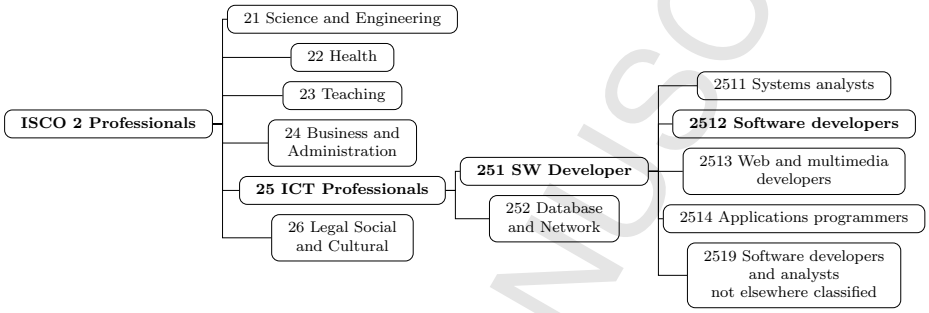

Figure 2: A branch of the ISCO classification tree for Professionals.

Definition 2 (**Job Occupation**). *A Job Occupation o is a tuple $O = (c, n)$ where $c \in \mathbb{N}$ is the occupation identifier, n is the text name of the occupation identified by c.*

To date, ISCO is the most adopted model worldwide at the basis of several national and international occupation classifiers. This is the case of the European classifier ESCO: the multilingual classification system of European Skills, Competences, Qualifications and Occupations. It is an ongoing project part of the Europe H2020 strategy, that is emerging as the European standard for supporting the whole labour market intelligence over 24 EU languages. Basically, the ESCO data model includes the ISCO hierarchical structure as a whole, and extends it through (i) a further level of fine-grained occupation descriptions and (ii) a taxonomy of skills, competences and qualifications. In our context we only focus on *skills*, we can formalise skills as follows.

Definition 3 (**skill**). *A skill s is a tuple $s = (us, ls)$ where us is the unique uri of the skill described by the text label ls.*

Text categorization aims at assigning a Boolean value to each pair $(d_j, c_i) \in D \times C$ where D is a set of documents and C a set of predefined categories. A *true* value assigned to (d_j, c_i) indicates document d_j to be set under the category c_i, while a false value indicates d_j cannot be assigned

under c_i. In our scenario, we consider a set of job vacancies $\mathcal{J}$ as a collection of documents each of which has to be assigned to one (and only one) ISCO occupation code. We can model this problem as a text classification problem, relying on the definition of [5].

Formally speaking, let $\mathcal{J} = \{J_1, \ldots, J_n\}$ be a set of Job vacancy according to Def.1, the classification of $\mathcal{J}$ under the a set of occupation codes O consists of |O| independent problems of classifying each job vacancy $J \in \mathcal{J}$ under a given ESCO occupation code o_i for $i = 1, \ldots, |O|$. Then, a *classifier* for o_i is a function $\psi : \mathcal{J} \times O \to \{0, 1\}$ that approximates an unknown target function $\dot{\psi} : \mathcal{J} \times O \to \{0, 1\}$. Clearly, as we deal with a single-label classifier, $\forall j \in \mathcal{J}$ the following constraint must hold: $\sum_{o \in O} \psi(j, o) = 1$.

Intuitively, once a job vacancy has been correctly classified on the lower level of the ISCO hierarchy (i.e., the fourth), one can navigate the ISCO/ESCO linkage accessing to a list of occupation examples and to the set of skills that characterise the corresponding occupation. This motivates the use of both ISCO and ESCO as target classification systems.

Definition 4 (**ESCO Classification System**). *The ESCO classification system $\mathcal{E}$ is a triple $\mathcal{E} = (O, R, S)$ where $O = \{o_1, \ldots, o_n\}$ is a set of job occupations, S is a set of skills $\{s_1, \ldots, s_m\}$, and $R : O \times S \to \mathbb{B}$ is the relation that associates a job occupation o to a skill s, namely $r(o, s) = 1$ iff the skills s is associated to the occupation o in ESCO.*

3.1. Feature Extraction

A Bag of Word Feature Extraction was applied to process job vacancy titles. Notice that this feature extraction pipeline is language independent as each step can be initiated by specifying the target language. Titles were pre-processed according to the following steps: (*i*) html tag removal, (*ii*) html entities and symbol replacement, (*iii*) tokenization, (*iv*) lower case reduction, (*v*) stop words removal (using the stop-words list provided by the NLTK framework [24]), (*vi*) stemming (using the Snowball stemmer), (*vii*) n-grams frequency computation (actually, unigram and bigram frequencies were computed, n-grams which appear less than 4 times or that appear in more than 30% of the documents are discarded, since they are not significant for classification). Each title is pre-processed according to the previous steps and is transformed into a set of n-gram frequencies.

3.2. Building up the Machine Learning Model

We built a machine learning model for classifying multilingual Web job vacancies exploiting a *single-label classifier* using both titles and descriptions. Indeed, titles often does not contain enough information for performing a correct classification (see, e.g., [25]). Several machine learning techniques have been evaluated for developing the text classifier, and they have been comparatively evaluated on a data set of 75,546 job vacancies in Italian, labelled by domain experts according to the most suitable four digits ISCO code. The techniques evaluated were: Support Vector Machines (SVMs), in particular SVM Linear [26], SVM RBF Kernel [27] , Random Forests (RFs) [28], and Artificial Neural Networks (ANNs) [29].

SVM classifiers have good generalisation ability; moreover, according to [30], they are well suited to the particular characteristics of texts, namely high dimensional feature spaces, few irrelevant features (dense concept vector), and sparse instance vectors.

Table 2: Classifiers Evaluation of the classifiers trained on the train set and evaluated on the test set. Precision, Recall, and F1-Score values are the weighted average of the values computed for each ISCO code.

Classifier	Precision	Recall	F1-Score
SVM Linear	**0.93**	**0.93**	**0.93**
SVM RBF Kernel	0.90	0.88	0.88
Random Forest	0.67	0.67	0.67
Neural Networks	0.92	0.92	0.92

Focusing on the Italian Labour Market Information, a set of 57,740 vacancies previously classified by domain experts belonging to ENRLMM[9] was used. The set of already classified vacancies was split into *train*, *validation*, and *test* sets. Then, a grid-search was performed over each classifier parameter space to identify the values maximizing the classification effectiveness (using training and validation sets). Tab. 2 shows the scores computed over the test set.

3.3. Evaluation

We used box plots to evaluate the f1-score measure of each classification algorithm over the first-digit of the ISCO taxonomy, which identifies 9 dis-

[9]The European Network on Regional Labour Market Monitoring

tinct occupation groups, from 1 up to 9. Box plot is a well-known statistical technique used in exploratory data analysis to visually identify patterns that may otherwise be hidden in a data set by measuring variation changes between different groups of data. In Figure 3 we report four box-plots, one for each classification algorithm. Each box-plot shows the distribution of the f1-score value for the respective algorithm over the nine ISCO groups. This allows us to focus on the effectiveness of each classification algorithm over a specific group of occupations.

Considering the f1-score measure, each distribution is partitioned into quartiles as follows: the *box* indicates the positions of the upper and lower quartiles respectively[10]; the interior of this box indicates the median value, which is the area between the upper and lower quartiles and consists of 50% of the distribution. Vertical lines (also known as *whiskers*) stretch over the extremes of the distribution indicating either minimum and maximum values in the dataset. Finally, dots are used to represent upper and lower outliers, namely data items that lie more (less) than 3/2 times the upper (lower) quartile respectively. As it can be observed, SVM Linear still outperforms the other classifiers even considering the overall ISCO distributions at the first level of the hierarchy. Furthermore, SVM Linear obtains a 0.91 median value for the f1-score with respect to the same value obtained by NN (0.88), RF (0.78), and SVM-RBF (0.7).

[10]The lower quartile is the 25th percentile while the upper quartile is the 75th percentile

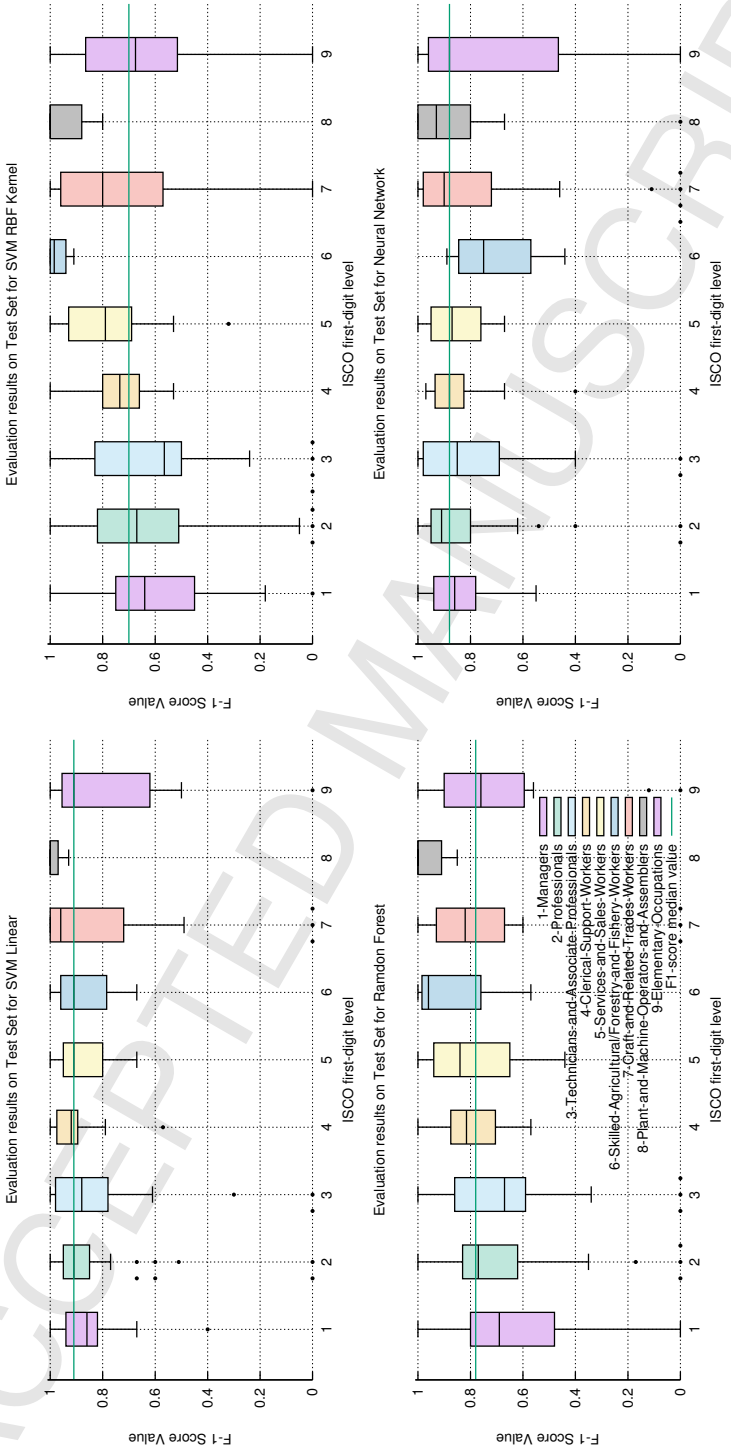

Figure 3: Evaluation results on Test Set for the machine-learning algorithms employed.

As it emerges from the described evaluation, by using machine learning techniques it is possible to achieve good performances in classifying Italian Web job vacancies with respect to the codes of the 4^{th} level of the ISCO taxonomy. Based on the evaluation results, we decided to employ the SVM linear classifier.

3.4. Skills Extraction

The text pieces stating the required skills usually concentrate on a small portion of the job vacancy descriptions. Thus, these relevant text pieces are extracted through a look-up search of sentinel expressions selected by domain experts involved within the projects. The domain expert kept adding new sentinel expressions until the number of n-grams identified in the next step grew (i.e., the set converged to a stable set). Then, the n-gram *Document Frequency* (DF), i.e., the number of vacancies where the n-gram is found, was computed considering as a scope both (1) the whole dataset and (2) the vacancy subsets homogeneous with respect to the ISCO occupation code.

The n-gram produced by the previous step underwent a string similarity comparison with respect to ESCO skill concept labels. The pair <*skill candidate* n-gram, ESCO Skill label> matching the following criteria were suggested for domain experts approval. The string similarity was computed as the mean value among the following well-known string metrics: Levenshtein distance, Jaccard similarity, and the Sørensen-Dice indexes[11]. The pairs having a similarity lower than 70% were dropped while the others were proposed to the domain experts for evaluation. Each <*candidate* n-gram, ESCO Skill label> above the threshold was reviewed by a domain expert to decide whether to consider an n-gram as: (1) a skill described in ESCO; (2) a skill not enlisted in ESCO (i.e., a *novel* skill); (3) or to reject the proposed n-gram as a skill concept. The outcome of this process can be seen as a dictionary of n-gram related skills and their corresponding ESCO skill concept (when available). Finally, the n-gram dictionary produced by the previous steps and the mappings among the ESCO skill concepts was used to look for skills among the downloaded vacancies.

[11]Though the latter is not a proper distance metric, it has been selected as we do not ask string metrics to satisfy the triangle inequality

3.5. Expressing Resulting Knowledge as a Graph

Thanks to the proposed approach we can *extend* the ESCO skills taxonomy through the skills extracted from the job vacancies, as well as to *weight* both occupations and skills with respect to the frequency through with which they appear in the Web Labour Market. The outcome is the knowledge graph that can be formalised as follows.

Definition 5 (**LMI Knowledge Graph**). *Let $\mathcal{E} = (O, R, S)$ be the ESCO classification system, and let $\psi : \mathcal{J} \times O \rightarrow \{T, F\}$ be a classification function, the LMI knowledge graph is a tuple $\mathcal{W} = (O, \bar{R}, \bar{S}, W_o, W_s)$ where:*
– $\bar{S} = S \cup S_{new}$ is the extended set of skills, where S_{new} is the set of skills extracted from the vacancies that can be reconciled to none of ESCO skills;
– $\bar{R} = R \cup R_{new}$ is the relation function that assigns a new skill $s \in S_{new}$ to the occupation o if and only if exists a job vacancy J such that $\psi(J, o) = 1$;
– W_o is a weight function that assigns to each occupation a real number, that is $W_o : O \rightarrow \mathbb{R}^+$ representing the share of Web job vacancies classified according to that occupation over the whole set of vacancies;
– W_s is a weight function that assigns to each occupation-skill relation $r(o, s) \in \bar{R}$ a real number, that is $W_s : \bar{R} \rightarrow \mathbb{R}^+$ representing the share of vacancies requiring the skill s over all the vacancies classified as occupation o.

The LMI Knowledge Graph would give a conceptual representation of the LMI knowledge base. The graph-based formalisation has been preferred here, as it allows a straightforward implementation through graph-databases.

The resulting knowledge on LM was then modelled as a graph. In Fig. 4 we report *a selection of* the graph-db data model according to the Neo4j property graph structure. The model is basically composed of two main node labels, *occupations* and *skills*. The former are the ISCO occupation codes whilst the latter are the union of both ESCO skills and the skills recognised as *novel* in the skill extraction phase, as described above. Then, two distinct directed relationships are allowed between skills and occupations to model that a skill s belongs to a given occupation o. The :BELONG relation would represent an occupation o requiring an ESCO skill s with a relevance of w in the ESCO taxonomy. This relationship measures the importance of the skill for a given occupation according to a set of labour market experts. By contrasty, the :BELONG_DATA relation models that w job vacancies have asked for skill s in the Web job vacancy text.

Such a knowledge graph allowed us to perform a several path-traversal analyses, such as *Skill2Job*, to identify the most promising occupations that one could be interested into given a set of skills, the *Gap Analysis* to identify the most important skills that one should acquire given a set of skills that a candidate already holds. Due to the space restrictions, here we only give the idea of how the graph-structure can be used to identify occupation groups on the basis of the skills they have in common, distinguishing between groups according to the *ESCO* and *real data*. To this end, we employed a local-clustering-coefficient metric to identify all the occupations that share *at least k%* skills in common (i.e., having a clustering coefficient equals to 1). We employed a weighted Jaccard metric to compute the similarity between occupations on the basis of skills in common.

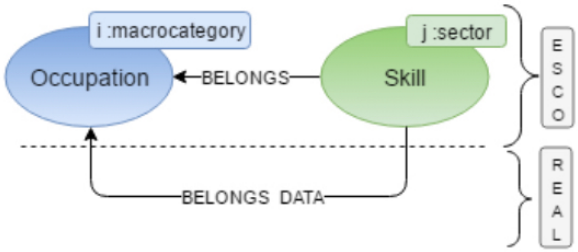

Figure 4: Extraction of the Graph-DB data model employed

Fig. 5 shows the ESCO skills category asked for in a group of occupations related to *mathematicians and statisticians*. The outer circle refers to ESCO skills whose relevance has been computed using the :BELONGS relation, whilst the inner circle refers to ESCO skills with a relevance computed using the :BELONG_DATA relation (i.e., Web job vacancies). As one might note, *computing* skills account for 6% according to the ESCO experts, whilst this value grows up to 66% in the real data that mainly specifies skills such as SQL, Relational Databases, Python and Data Warehouse. Conversely, the *Business & Administration* sector seems to be overestimated by the ESCO taxonomy, which indicates up to 56 B&A related skills whilst only few on them are in fact actually repeatedly asked for companies. This analysis would allow one to measure the gap between (1) an LMI system built through an expert-driven approach as in the ESCO case, where labour market experts

indicates a list of skills that are relevant in an occupation profile, and (2) a data-driven system, where skills are recognised as important on the basis of the real labour market expectations.

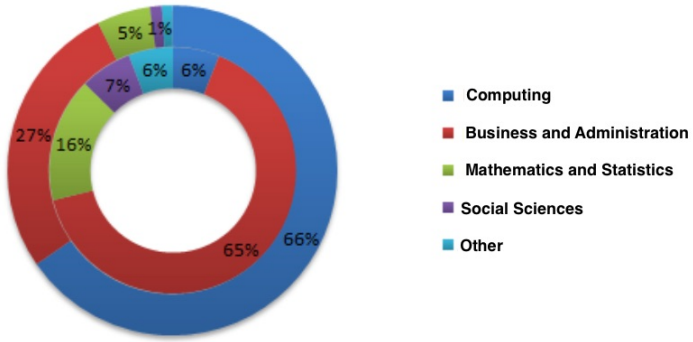

Figure 5: It includes the following ESCO occupations: *Statistical, mathematical and related associate professionals*, *Financial Analysts* and *Mathematicians, actuaries and statisticians* having at least 60% skills in common.

4. Results

A crucial step in the LMI activity here described relies on the vacancy collection and dataset preparation. This task has been performed by crawling the Web, selecting a set of vacancies from those collected, and by labelling each item with the most appropriate 4th level ISCO code. The benchmark dataset was built with the support of a team of labour experts belonging to the ENRLMM network[12]. Though the data collected (and used) for WollyBI and Cedefop projects are quite different, the collection and dataset preparation phase followed a similar pipeline:

- The team of experts identified the Web sites playing a prominent role in the domestic job market, for each country involved in each project;
- Each selected Web site was crawled at least once per week, by identifying and downloading the fresh job vacancies;

[12]The European network on regional labour market monitoring

- A set of vacancies were selected and labelled by the domain experts; this reduced dataset was then used to perform training, test, and validation of the machine learning based classifiers

The WollyBI Project. WollyBI is a SaaS tool for collecting and classifying Web job vacancies on the ISCO/ESCO standard taxonomies, and extracting the most requested skills from job descriptions. It has been designed to provide five distinct entry points to the user on the basis of the analysis purposes, which are, *Geographical Area*, *Skill*, *Firm*, *Occupation*, and *free OLAP queries*.

- *Geographical Area* to find the most searched occupations on the Web and their skills at a very fine-grained geographical level;
- *Skill* to input a bag-of-skills, and to discover the most searched occupations that include the given skills (i.e., profile gap-analysis);
- *Firm* to obtain a rank of occupations that specifies a certain industry sector in the vacancy;
- *Occupation* to navigate through the ISCO/ESCO classifications and to exploit the details related to each occupation;
- *Customised* to perform classical drill-down and roll-up operations freely over the OLAP cubes.

On-line Demo. A demo video on the Geographical Area has been made available at `https://goo.gl/4cokib` while a demo video of WollyBI in action on the Occupation dimension has been made available at `https://youtu.be/zBNsAS5L04g`.

4.1. Real-life Study: Competition Analysis for Strategic Decision Making

WollyBI supported a recruitment agency in identifying and measuring the market share with respect to its competitors, including the most important recruitment agencies in Italy, namely: GiGroup, Adecco, ManPower, RandStad, ObiettivoLavoro, and Umana. In Fig. 8 we report the market share distribution over the top-10 ISCO occupations, by analysing the Italian Web Job Vacancies between February 2013 and April 2015. Clearly, due to non-disclosure agreements, the agency labels reported in Fig. 8 have been

anonymised. We analysed about 850K Web job vacancies posted by these agencies.

This competition analysis has been validated as helpful by the customer (Agency B) as it allowed one (i) to *measure* its position in the market and the *gap* with respect to their competitors; (ii) to *drive* the identification of strategic decisions to improve its market share and, in turn, to identify the corresponding strategies that allow achievement of the desired goals. Just to give a few examples, our analysis revealed that Agency B is leader in recruiting "shop sales assistants" whilst its market share ranges between 9% and 15% in the remaining professions. Thanks to these results, Agency B has been enabled to design its strategic intervention through fact-based decision-making.

4.2. Real-life study: Educational Training

This study aimed at analysing both job-specific and traversal skills for a set of selected occupational categories. This study is part of the "University and Companies 2016" report by CRUI (The Conference of Italian University Rectors)[13] which aims to foster the cooperation of universities and industries by reducing the mismatching between the labour market and research. The purpose of our study was to identify the most requested skills by occupation categories, using the ESCO skill taxonomy as a baseline for sharing and representing our results through a well-recognised international taxonomy of skills. To this end, the ISCO occupations have been grouped on 5 distinct areas, focusing only on occupations that require a master degree: *Administrative and Management*, *General Direction*, *Marketing and Sales*, *Production and Logistics* and *Information and Communication Technologies (ICT)*. Notice that these areas cover over 260K+ Web job vacancies collected by WollyBI from 2013 until May 2016. Fig. 6 shows the job-specific skills distribution over the five areas we identified. Not surprisingly, skills related to Business and Administrations outperforms the others in the first three areas. Conversely, they have a lesser impact on "Production and Logistic" and "ICT" areas, which are dominated by "computing" skills.

The results obtained have been validated as helpful for improving educational policies with the aim if strengthening the relation between universities and enterprises.

[13] `https://www.crui.it/crui-english.html` - executive summary available at `http://www.universitaimprese.it/report-annuale-2016/`

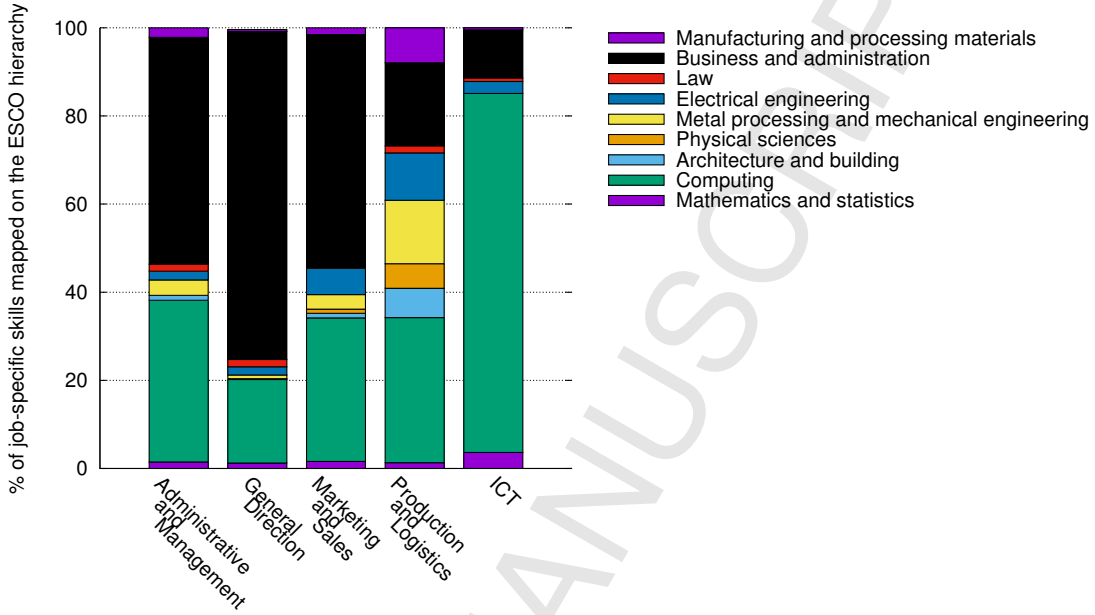

Figure 6: The distribution of the ESCO job-specific skills for each Area analysed. The whole ESCO skill hierarchy can be accessed at `https://goo.gl/QFSjGh`

4.3. EU Project: The Cedefop Experience

In 2014, our experience of WollyBI was the basis of a prototype system that we realised within a call-for-tender for the Cedefop EU Agency: thi was aimed at collecting Web job vacancies from five EU countries and extracting the requested skills from the data. The rationale behind the project was to turn data extracted from Web job vacancies into knowledge (and thus value) for policy design and evaluation through fact-based decision making. The system is running on the Cedefop data centre since June 2016, gathering and classifying job vacancies from 5 EU countries, namely: United Kingdom, Ireland, Czech Republic, Italy and Germany. To date, the system has collected 7+ million job vacancies over the 5 EU countries, and it accounts among the research projects that a selection of Italian universities addressed in the context of big data [31]. In Fig. 7 we report a snapshot from the project dashboard that allows to surf the data over the ISCO taxonomy and the ESCO skills extracted from the data.

This prototype enabled the Cedefop agency to have evidence of the *competitive advantage* that the analysis of Web job vacancy permits compared to the classical survey-based analyses, in terms of (i) near-real time labour

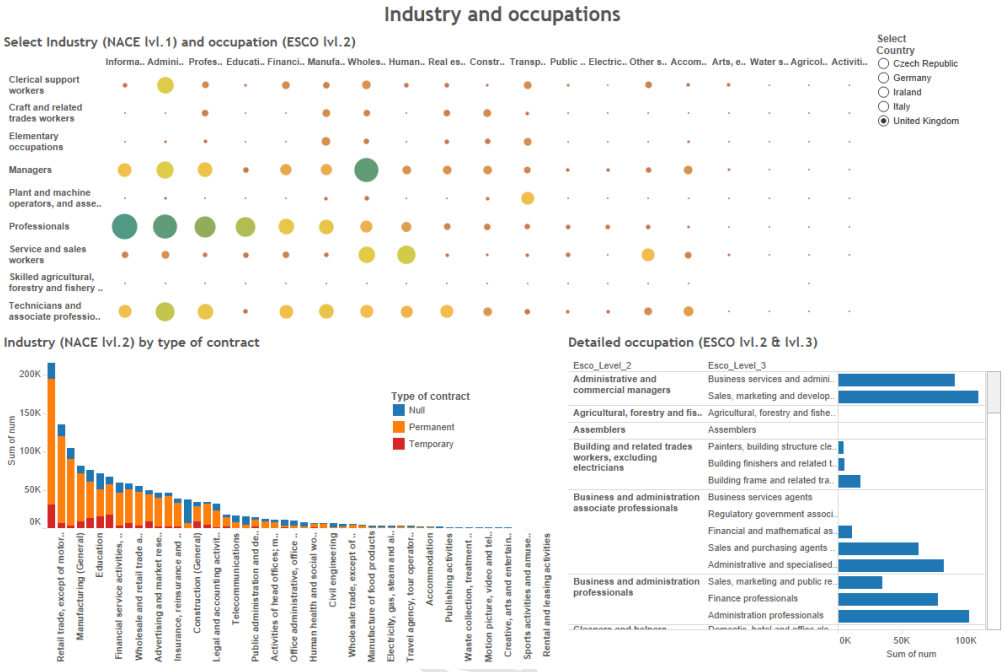

Figure 7: A snapshot from the System Dashboard deployed for CEDEFOP Project. Interactive Demo available at https://goo.gl/bdqMkz

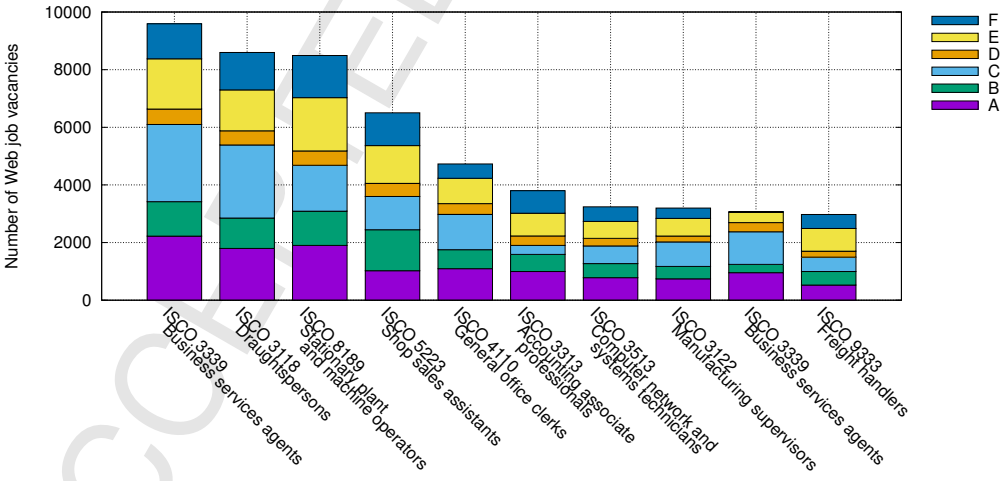

Figure 8: Top-10 ISCO occupations and relative distribution over the six recruitment agencies analysed.

market information; (ii) reduced time-to-market; (iii) fine-grained territorial analysis, from countries down to municipalities; (iv) identification of the most requested skills, and (v) evaluation of skills impact on each ISCO profession. *On-line Demo.* An animated heat-map showing the job vacancy collection progress over time is available at `http://goo.gl/MQUiUn`. Finally, a demo video of the system prototype Web interface is available at `http://goo.gl/RRb63S`.

5. Concluding Remarks and Research Directions

In this paper we have described our approach to Web Labour Market Intelligence along with three real-life application scenarios, focusing on the realisation of a machine learning model for classifying job vacancies. The main benefits of our approach to LMI are: (i) reduced the time-to-market with respect to classical survey-based analyses; (ii) multi language support through the use of standard classification systems - rather than proprietary ones - by overcoming linguistic boundaries over countries; (iii) the ability to represent the resulting knowledge over several dimensions (e.g., territory, sectors, contracts, etc.) at different level of granularity, and (iv) the ability to evaluate and compare international labour markets to support fact-based decision making.

Our research goes in two directions. From an application point of view, we have been engaged by Cedefop to extend the prototype to the whole EU community to all 28 EU Member States, building the system for the EU Web Labour Market Monitoring[14].

From a methodological perspective, reasoning with Web job vacancies raises some interesting research issues, such as the automatic synthesis of the labour market knowledge through word embeddings, the identification of AI heuristic-search algorithms for path-traversal over big knowledge-graph, as well as the design of novel AI techniques for data cleansing in a big data scenario.

We are actually working on applying word-embedding to our labour knowledge graph, as this would allow representing lexicon differences in the different Countries. Furthermore, we are working on extending a prelimi-

[14] "Real-time Labour Market information on Skill Requirements: Setting up the EU system for online vacancy analysis AO/DSL/VKVET-GRUSSO/Real-time LMI 2/009/16. Contract notice - 2016/S 134-240996 of 14/07/2016 `https://goo.gl/5FZS3E`

nary approach we used to identify new potential occupations [32] through language models. Finally, we are also working with economists at CRISP Research Centre for investigating prediction analytics and data mining algorithms to forecast the job vacancies and skill demand by employers based on historical data.

References

[1] I. Lee, Modeling the benefit of e-recruiting process integration, Decision Support Systems 51 (1) (2011) 230–239.

[2] A. Singh, C. Rose, K. Visweswariah, V. Chenthamarakshan, N. Kambhatla, PROSPECT: a system for screening candidates for recruitment, in: Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, 659–668, 2010.

[3] X. Yi, J. Allan, W. B. Croft, Matching resumes and jobs based on relevance models, in: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 809–810, 2007.

[4] W. Hong, S. Zheng, H. Wang, Dynamic user profile-based job recommender system, in: Computer Science & Education (ICCSE), IEEE, 2013.

[5] F. Sebastiani, Machine learning in automated text categorization, ACM computing surveys (CSUR) 34 (1) (2002) 1–47.

[6] F. H. Khan, S. Bashir, U. Qamar, TOM: Twitter opinion mining framework using hybrid classification scheme, Decision Support Systems 57 (2014) 245 – 257.

[7] A. Zubiaga, D. Spina, R. Martínez-Unanue, V. Fresno, Real-time classification of Twitter trends, JASIST 66 (3) (2015) 462–473.

[8] P. Melville, W. Gryc, R. D. Lawrence, Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2009.

[9] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in: ACL-02 conference on Empirical methods in natural language processing, Association for Computational Linguistics, 2002.

[10] C.-H. Chang, M. Kayed, M. R. Girgis, K. F. Shaalan, A Survey of Web Information Extraction Systems, IEEE Transactions on Knowledge and Data Engineering 18 (10) (2006) 1411–1428.

[11] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni, Open Information Extraction from the Web., in: IJCAI, vol. 7, 2670–2676, 2007.

[12] E. F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition, in: Conference on Natural Language Learning at HLT-NAACL, 2003.

[13] M. Zhao, F. Javed, F. Jacob, M. McNair, SKILL: A System for Skill Identification and Normalization., in: In the Twenty-Seventh AAAI Conference on Innovative Applications of Artificial Intelligence, AAAI, 4012–4018, 2015.

[14] K. Yu, G. Guan, M. Zhou, Resume information extraction with cascaded hybrid model, in: In the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 499–506, 2005.

[15] A. De Sitter, W. Daelemans, Information extraction via double classification, in: International Workshop on Adaptive Text Extraction and Mining, 2003.

[16] I. Kivimäki, A. Panchenko, A. Dessy, D. Verdegem, P. Francq, C. Fairon, H. Bersini, M. Saerens, A graph-based approach to skill extraction from text, in: Workshop on Graph-based Methods for Natural Language Processing, 2013.

[17] B. Liu, Web data mining: exploring hyperlinks, contents, and usage data, Springer Series on Data-Centric Systems and Applications, Springer, doi:\bibinfo{doi}{10.1007/978-3-642-19460-3}, URL `http://www.springer.com/gp/book/9783642194597`, 2007.

[18] J. Han, J. Pei, M. Kamber, Data mining: concepts and techniques, Elsevier, 2011.

[19] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, 2016.

[20] M. Mezzanzanica, R. Boselli, M. Cesarini, F. Mercorio, A model-based evaluation of data quality activities in KDD, Information Processing & Management 51 (2) (2015) 144–166.

[21] T. Dasu, Data Glitches: Monsters in Your Data, in: Handbook of Data Quality, Springer, 163–178, 2013.

[22] A. Haug, F. Zachariassen, D. Van Liempd, The costs of poor data quality, Journal of Industrial Engineering and Management 4 (2).

[23] R. Boselli, M. Mezzanzanica, M. Cesarini, F. Mercorio, Planning meets Data Cleansing, in: The 24th International Conference on Automated Planning and Scheduling (ICAPS 2014), AAAI, 439–443, 2014.

[24] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, " O'Reilly Media, Inc.", 2009.

[25] F. Amato, R. Boselli, M. Cesarini, F. Mercorio, M. Mezzanzanica, V. Moscato, F. Persia, A. Picariello, Challenge: Processing web texts for classifying job offers, in: IEEE International Conference on Semantic Computing, 2015.

[26] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: A library for large linear classification, The Journal of Machine Learning Research 9 (Aug) (2008) 1871–1874.

[27] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf, An introduction to kernel-based learning algorithms, Neural Networks, IEEE Transactions on 12 (2) (2001) 181–201.

[28] L. Breiman, Random Forests, Machine Learning 45 (1) (2001) 5–32.

[29] S. Haykin, N. Network, A comprehensive foundation, Neural Networks 2 (2004) (2004) 41.

[30] T. Joachims, Text categorization with Support Vector Machines: Learning with many relevant features, in: C. Nédellec, C. Rouveirol (Eds.), Machine Learning: ECML-98, vol. 1398 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, ISBN 978-3-540-64417-0, 137–142, doi:\bibinfo{doi}{10.1007/BFb0026683}, URL http://dx.doi.org/10.1007/BFb0026683, 1998.

[31] S. Bergamaschi, E. Carlini, M. Ceci, B. Furletti, F. Giannotti, D. Malerba, M. Mezzanzanica, A. Monreale, G. Pasi, D. Pedreschi, et al., Big Data Research in Italy: A Perspective, Engineering 2 (2) (2016) 163–170.

[32] S. Marrara, G. Pasi, M. Viviani, M. Cesarini, F. Mercorio, M. Mezzanzanica, M. Pappagallo, A Language Modelling Approach for Discovering Novel Labour Market Occupations from the Web, in: 2017 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2017), ISBN 978-1-4503-4951-2, 1026–1034, doi:\bibinfo{doi}{10.1145/3106426.3109035}, URL http://doi.acm.org/10.1145/3106426.3109035, 2017.

Roberto Boselli is Assistant Professor in Computer Science at the University of Milan - Bicocca, Department of Statistics and Quantitative Methods. He is also affiliated to the CRISP research centre. He worked in several national and international research projects in the field of information systems and methodologies for semantic processing of information. His research activities focus on Semantic Web, Social Media Analytics, Information Extraction and e-government services. He published many scientific papers in books, international journals, proceedings of national and international conferences.

Mirko Cesarini's Profile

Mirko Cesarini is an Assistant Professor in Computer Science at the University of Milan Bicocca, department of Statistics and Quantitative Methods. He was awarded a PH.D. in Computer Science from Politecnico di Milano and has published several papers in international journals and conference proceedings. He is in the program committee of several international journals and conference.

His research activities focus on Data Quality, Natural Language Processing, Text Classification, Machine Learning, and Information Systems.

He has worked on several international research projects on Data Management and Knowledge Discovery, information extraction from the web for statistical purpouses, and on Big Data Analytics .

Fabio Mercorio is Assistant Professor of Computer Science at University of Milan-Bicocca, dept. of Statistics and Quantitative Methods. He holds a PhD in Computer Science and Application in 2012 at the University of L'Aquila, Italy. Since 2011 he has been research fellow at CRISP. His research interests include Artificial Intelligence Planning and Knowledge Discovery. His main contribution is in the design and development of an automated algorithm for performing Data Quality Analysis and Cleansing tasks through model-checking. On the AI Planning side, he contributed as co-designer and co-developer of UPMurphi, the state-of-the-art planner for hybrid systems specified through the PDDL+ language. He collaborated in the realisation of DiNo, an heuristic planner based on UPMurphi developed by the AI Planning Group at King's College London. As member of the CRISP research team, he has been involved in many national and International research projects related to Data Management and Knowledge Discovery for supporting the decision making activities.

Mario Mezzanzanica is Associate Professor of "Information Systems" at the University of Milan Bicocca. He is the Scientific Director of the "CRISP" centre - Inter University Research Centre on public services. His research interests include Information Systems, Databases, Artificial Intelligence, Business Intelligence and Knowledge Discovery. He is member of the editorial board of international journals in the field of Artificial Intelligence and Information Systems. He has also been involved in several technical and scientific committees activated by Public Institutions, aimed at studying new models and methodologies to design, monitor and evaluation of innovation projects, with relevant impacts on ICT-based public services. He has a strong experience in applying his research to real-life scenarios, as he partnered with the role of advice and coordination in several projects for innovation of public services at both national and regional level, working with the Italian Presidency of the Council of Ministers, the Italian Ministry of Labour and Social Security, the ISTAT (Italian National Institute for Statistics), the AIPA (Authority for Information Technology in Italian Public Administration), the Italian Ministry of Economy and Finance, CONSIP (Central Public Procurement Agency, a public stock company owned by Ministry of the Economy and Finance), as well as the main Italian Regions as Lombardy, Veneto, and Emilia Romagna. Since 2010 he is the promoter and director of the Master "Business Intelligence and Decision Support Systems" provided by University of Milan Bicocca.

Highlights

— Classification of Job advertisements collected from Web-sites is becoming a relevant issue to both academic and industry communities, as it allows analysing, observing, and comparing how real labour market dynamics evolve over countries;

— we build a machine learning model for classifying multilingual Web job vacancies, fully implemented into a system and an EU research project;— we report experimental evaluation of machine learning algorithms employed in two real-life scenarios of Web Labour Market along with the contributions provided to support the decision making activities of end-users;— we show how the knowledge base on the Web labour market can be modelled as a graph to allow for graph-traversal queries.