Labour Market Intelligence for Supporting Decision Making

Roberto Boselli[1,2], Mirko Cesarini[1,2], Fabio Mercorio[1,2], and Mario Mezzanzanica[1,2]

[1] Dept. of Statistics and Quantitative Methods, Univ. of Milano-Bicocca, Italy
[2] CRISP Research Centre, Univ. of Milano-Bicocca, Italy

(discussion paper)

Abstract. Over the past decade, a growing number of employers has been using the Web for advertising job opportunities through Web job vacancies, that usually specify a job position, along with a set of skills a candidate should possess. Reasoning with these Web job advertisements can effectively support the decision marking processes of several labour market stakeholders, including public organisations, educational and employment agencies, and analysts as well. Here, Labour Market Intelligence refers to the design and definition of automated methodologies and tools for supporting real-time labour market monitoring at a very fine-grained level. This, in turn, represents a competitive advantage to labour market stakeholders with respect to classical survey-based analyses, as they are quite expensive and may require up to one year before being available. In this paper we discuss how Web job vacancies have been collected from selected websites, processed, and classified over a standard taxonomy through machine learning algorithms, extracting the most relevant skills from raw texts. Then, we show how our approach has been applied to some real-life studies and we discuss the benefits provided to end users.

Keywords: Machine Learning, Text Classification, Big Data

1 Introduction

In recent years, the diffusion of Web-centric services is growing exponentially, and this allows a significant part of European Labour demand to be conveyed through specialised Web portals and services. This also contributed to introduce the term ”Labour Market Intelligence” (LMI), that refers to the use and design of AI algorithms and frameworks to Labour Market Data for supporting decision making. This is the case of *Web job vacancies*, that are job advertisements containing two main text fields: a *title* and a *full description.* The title shortly summarises the job position, while the full description field usually includes the position details and the relevant skills the employee should hold.

There is a growing interest in designing and implementing real LMI application to Web Labour Market data for supporting the policy design and evaluation activities through evidence-based decision-making. In 2016, the European Commission's 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 extracting 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”.

The rationale behind all these initiatives is that reasoning over Web job vacancies represents an added value for both *public and private* labour market operators to deeply understand the Labour Market dynamics, occupations, skills, and trends: (*i*) by reducing the time-to-market with respect to classical survey-based analyses (official Labour Market surveys results actually require up to one year before being available); (*ii*) by overcoming the linguistic boundaries through the use of standard classification systems rather than proprietary ones; (*iii*) by representing the resulting knowledge over several dimensions (e.g., territory, sectors, contracts, etc) at different level of granularity and (*iv*) by evaluating and comparing international labour markets to support fact-based decision making.

Paper’s Goal. This paper would summarise some results of our research activities and outcomes in Web Labour Market Intelligence in two distinct research projects: *WollyBI*[4] and *Cedefop*[3], by focusing on (i) job vacancy classification and skill extraction tasks, and (ii) the support to decision making activities that our approach provided to the involved stakeholders.

2 Web Labour Market in the Literature

LMI is an*emerging* cross-disciplinary field of studies that is attracting research interests in both industrial and academic communities, as we summarise below.

Scientific 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” [12]. 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. Recently, text classification has proven to give good results in categorizing many real-life Web-based data such as, for instance, news and

[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

social media [7], and sentiment analysis [11]. On the other side, skills extraction from Web job vacancies can be framed in the Information Extraction field [6] and Named Entity Recognition [13]. The latter has been applied to solve numerous domain specific problems in the areas of Information Extraction and Normalization [14]. In the last years, public administrations started exploring new ways for supporting knowledge management (see, e.g., [2]) as well as for obtaining detailed and fresh information about the Labour Market. Here, administrative information collected by public administrations has been used for studying the Italian Labour Market dynamics performing both data quality [5,9] and knowledge discovery activities [10,4] through AI techniques (see, e.g.[8]). Unfortunately, administrative data are collected when people is hired (and only in countries where the state collect such information), therefore they do not provide information about the labour demand.

Industries. This problem is also relevant for business purposes, and this motivates the growing of several commercial products providing job seekers and companies with skill-matching tools. Concerning firms, they strongly need to automatize Human Resource (HR) department activities; as a consequence, a growing amount of commercial skill-matching products has been developed in the last years, for instance BurningGlass, Workday, Pluralsight, EmployInsight, and TextKernel. To date, the only commercial solution that uses standard taxonomies as thesauri is Janzz: a Web based platform to match 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 classyfying job vacancies through the Google Machine Learning service over O*NET, that 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.

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. Second, our approach aims to build a knowledge-graph for supporting the fact-based decision making activities for LMI.

3 An Approach to deal with Web Job Vacancies

Background on Web Labour Market Information. A Web job offer extracted 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.

One of the most important classification system designed for this purposes is ISCO: The *International Standard Classification of Occupations* is a four-level classification that represents a standardised way for organising the labour market occupations. ESCO is the multilingual classification system of European Skills, Competences, Qualifications and Occupations, it is the European standard for supporting the whole labour market intelligence over 24 EU languages. Basically, the ESCO data model includes the whole ISCO structure, and extends it through (i) a further level of fine-grained occupation descriptions and (ii) a taxonomy of skills, and competences.

Web job vacancy classification through Text Classification. 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 [12]. Formally speaking, let $\mathcal{J} = \{J_1, \ldots, J_n\}$ be a set of Job vacancies, the classification of $\mathcal{J}$ under $|O|$ ESCO occupation labels 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 $\psi : \mathcal{J} \times O \rightarrow \{0, 1\}$ that approximates an unknown target function $\dot{\psi} : \mathcal{J} \times O \rightarrow \{0, 1\}$. Clearly, as we deal with a single-label classifier, $\forall j \in \mathcal{J}$ the following constraint must hold: $\sum_{o \in O} \psi(j, o) = 1$. Notice that in this way we can also *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 which they appear in the Web Labour Market.

Building up the Machine Learning Model. We build 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 as we shown in [1]. 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. The evaluated techniques are: Support Vector Machines (SVMs), in particular SVM Linear, SVM RBF Kernel, Random Forests (RFs), and Artificial Neural Networks (ANNs).

Table 1: 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[5] was used.

[5] The European Network on Regional Labour Market Monitoring

The set of already classified vacancies are 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. 1 shows the scores computed over the test set.

Skills Identification. The text pieces stating the required skills usually concentrate on a small portion of 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 grows (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, is computed considering as a scope both (1) the whole dataset and (2) the vacancy subsets homogeneous w.r.t. 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[6]. The pairs having a similarity lower than 70% are dropped while the others are proposed to the domain experts for evaluation. Each <*candidate* n-gram, ESCO Skill label> above the threshold has been reviewed by a domain expert to decided 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 have been used to look for skills among the downloaded vacancies.

LM Knowledge as a Graph. The resulting knowledge on LM is then modelled as a graph. In Fig. 1 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. Differently, 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

[6] 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

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 weigthed Jaccard metric to compute the similarity between occupations on the basis of skills in common. Fig. 2 shows the ESCO skills category asked for 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 ESCO taxonomy, that indicates up to 56 B&A related skills whilst only few on them are actually repeatedly asked by 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.

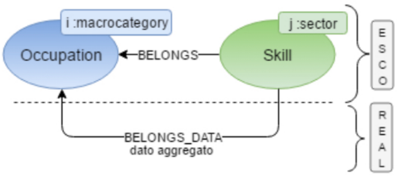

Fig. 1: Extraction of the Graph-DB data model employed

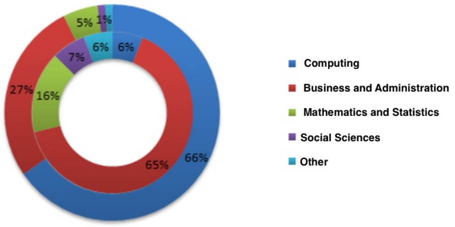

Fig. 2: It includes the following ESCO occupations: *Statistical, mathematical and related associate professionals*, *Financial Analysts* and *Mathematicians, actuaries and statisticians* having at least 60% skills in common.

4 Some Results and Concluding Remarks

The WollyBI Experience. 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, that are, *Geographical Area*, *Skill*, *Firm*, *Occupation*, and *free OLAP queries*.

Competition Analysis for Strategic Decision Making. WollyBI supported a recruitment agency to identify and measure the market share with respect to its competitors, that included the most relevant recruitment Agencies in Italy, namely: GiGroup, Adecco, ManPower, RandStad, ObiettivoLavoro, and Umana. In Fig. 4 we report the market share distribution over the top-10 ISCO occupations, by analysing the Italian Web Job Vacancies since February 2013 to April 2015. Clearly, due to undisclosure agreements, the agency labels reported in Fig. 4 have been anonymized. We analysed about 850K Web job vacancies posted by these agencies. This competition analysis has been validated as helpful to the customer (Agency B) as it allowed to (i) *measure* the 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 achieving 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 this results, Agency B has been made in state to design its strategic intervention through fact-based decision-making.

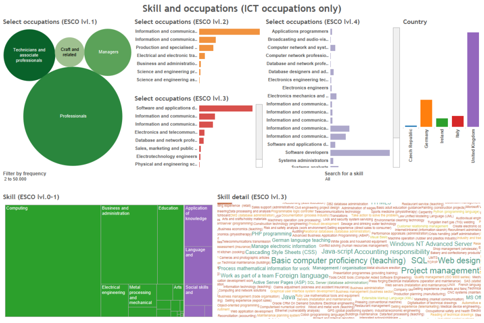
Skill and occupations (ICT occupations only)
Select occupations (ESCO lvl.1)
Technicians and associate professionals
Craft and related
Managers
Professionals
Select occupations (ESCO lvl.2)
Information and communica..
Information and communica..
Production and specialised ..
Electrical and electronic tra..
Business and administratio..
Science and engineering pr..
Science and engineering as..
Select occupations (ESCO lvl.3)
Software and applications d..
Information and communica..
Information and communica..
Electronics and telecommun..
Database and network profe..
Sales, marketing and public ..
Electrotechnology engineers
Physical and engineering sc..
Select occupations (ESCO lvl.4)
Applications programmers
Broadcasting and audio-vis..
Computer network and syst..
Computer network professio..
Database and network profe..
Database designers and ad..
Electronics engineering tec..
Electronics engineers
Electronics mechanics and ..
Information and communica..
Information and communica..
Information and communica..
Information and communica..
Information and communica..
Software and applications d..
Software developers
Systems administrators
Systems analysts
Country
Czech Republic
Germany
Ireland
Italy
United Kingdom
Filter by frequency
2 to 50.000
Search for a skill
All
Skill (ESCO lvl.0-1)
Computing
Business and administration
Education
Application of knowledge
Language and
Electrical engineering
Metal processing and mechanical
Arts
Social skills and
Skill detail (ESCO lvl.3)
PHP programming
German language teaching
Windows NT Advanced Server
Manage electronic information
Cascading Style Sheets (CSS)
Java-script
Accounting responsibility
Basic computer proficiency (teaching)
SQL
Web design
Process mathematical information for work
Management / organisation
Project management
Work as part of a team
Foreign language
Active Server Pages (ASP)
SQL Server (database administration)
Java
MS Office

Fig. 3: Snapshot from the prototype as deployed to the Cedefop Agency

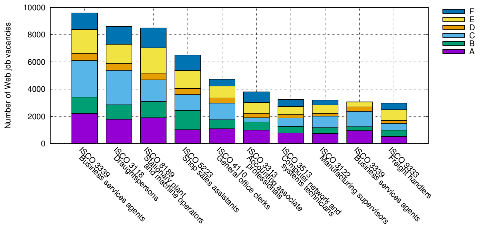

Fig. 4: Top-10 ISCO occupations and relative distribution over the six recruitment agencies analysed.

The Cedefop Experience. In 2014, the experience of WollyBI put the basis of a prototype system we realised within a call-for-tender for the Cedefop EU Agency aimed at collecting Web job vacancies from five EU countries and extracting 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 fact-based decision making. The architecture of the system basically relies on WollyBI, that is now 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 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 [3]. In Fig. 3 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.

Concluding Remarks and Research Directions. In this paper we described our approach to Web Labour Market Intelligence, focusing on the realisation of a machine learning model for classifying job vacancies and showing the benefits of a graph-based representation of the knowledge base. Our research goes toward two directions. From an application point of view, we have been commited by Cedefop for extending the prototype to the whole EU community to all 28 EU Countries, building the system for the EU Web Labour Market Monitoring[7]. 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 cleasing in a big data scenario.

References

1. Amato, F., Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., Moscato, V., Persia, F., Picariello, A.: Challenge: Processing web texts for classifying job offers. In: IEEE International Conference on Semantic Computing (2015)
2. Amato, F., Colace, F., Greco, L., Moscato, V., Picariello, A.: Semantic processing of multimedia data for e-government applications. J. Vis. Lang. Comput. 32 (2016)
3. Bergamaschi, S., Carlini, E., Ceci, M., Furletti, B., Giannotti, F., Malerba, D., Mezzanzanica, M., Monreale, A., Pasi, G., Pedreschi, D., et al.: Big data research in italy: A perspective. Engineering 2(2), 163–170 (2016)
4. Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M.: Inconsistency knowledge discovery for longitudinal data management: A model-based approach. In: SouthCHI13 special session on Human-Computer Interaction & Knowledge Discovery. Springer (2013)
5. Boselli, R., Mezzanzanica, M., Cesarini, M., Mercorio, F.: Planning meets data cleansing. In: The 24th International Conference on Automated Planning and Scheduling (ICAPS 2014). pp. 439–443. AAAI (2014)
6. Chang, C.H., Kayed, M., Girgis, M.R., Shaalan, K.F.: A survey of web information extraction systems. IEEE Transactions on Knowledge and Data Engineering 18(10), 1411–1428 (2006)
7. Khan, F.H., Bashir, S., Qamar, U.: Tom: Twitter opinion mining framework using hybrid classification scheme. Decision Support Systems 57, 245 – 257 (2014)
8. Mercorio, F.: Model checking for universal planning in deterministic and non-deterministic domains. AI Communications 26(2), 257–259 (2013)
9. Mezzanzanica, M., Boselli, R., Cesarini, M., Mercorio, F.: Data quality through model checking techniques. In: Intelligent Data Analysis (IDA), LNCS. pp. 270–281. Springer (2011)
10. Mezzanzanica, M., Boselli, R., Cesarini, M., Mercorio, F.: A model-based evaluation of data quality activities in KDD. Information Processing & Management 51(2), 144–166 (2015)
11. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: ACL-02 conference on Empirical methods in natural language processing. Association for Computational Linguistics (2002)
12. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) 34(1), 1–47 (2002)
13. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Conference on Natural Language Learning at HLT-NAACL (2003)
14. Zhao, M., Javed, F., Jacob, F., McNair, M.: Skill: A system for skill identification and normalization. In: In the Twenty-Seventh AAAI Conference on Innovative Applications of Artificial Intelligence. pp. 4012–4018. AAAI (2015)

[7] “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