Web Job Vacancies as a Resource for Better Labor Market Knowledge

2

by Silvia Dusi

Introduction

"Every day, 2.5 quintillion bytes of data are created. This data comes from digital pictures, videos, posts to social media sites, intelligent sensors, purchase transaction records, cell phone GPS signals to name a few." (Zicari 2014). The term "Big Data" usually refers to large amounts of different types of data produced with high velocity from a high number of various types of sources.

Over the last years, social media (the main Web 2.0 websites and services), and in particular the social networks, have been gaining more and more importance in Italy: over 60% of Italians who use the Internet have a Facebook profile (23 million out of 38), and 3.5 million have a LinkedIn account; compared to 2012, the number of LinkedIn profiles increased more than 18%. The number of Italian companies that use social media for professional reasons, such as a Facebook page or an account on LinkedIn or Twitter, also increased. According to a recent survey, almost 40% of Italian companies are already present on a social network, with a further 32% considering it (Osservatorio Business Intelligence, 2012).

In this context, social media are more and more becoming a vehicle for innovative and effective services for citizens, businesses, and governments; among these online services, an increasingly important role is played by those focusing on employment. Social recruiting (e.g., via the Jobvite website) is a new form of matching the supply and the demand of labor on the Web. It is made possible mainly by the growing opportunities social media offer for building relationships and facilitating communication flows. The Web 2.0 has definitely changed the way we seek new job opportunities, increased the collaboration between subjects, and expanded the channels of information dissemination. Moreover, the Web today can provide an important contribution to the labor market domain, as the huge amount of data available online constitutes a useful resource containing information on the trends and dynamics of the labor market under several points of view, such as geographical area, occupations, and skills.

We can say that the Web is becoming more and more popular to express both the Labor Demand and Supply because it allows:

- Better intermediation services between people and companies
- Decreasing the mismatch
- Production of information on actual trends to improve the stakeholders' policies

Clearly, there is a need to follow an appropriate methodology for processing the information derived from a large amount of unstructured text, such as Web Job vacancies that companies publish through specialized websites (Dalal and Zaveri 2013, Feldman and Sanger 2007, McCallum 2005).

The term "Big" does not refer only to the quantity of data, but also to the heterogeneity of data sources and to the velocity in analyzing data. In the literature, the 3Vs model describes the fundamental dimensions of Big Data as:

1. Volume: refers to the massive size of data generated by machines, networks, and human interaction within systems (e.g., social media);
2. Velocity: refers to the pace at which data are generated (e.g., real-time) leaving the delay of analysis very short;
3. Variety: data structure and contents, e.g., semi-structured or unstructured data with video, photo, web links.

The main challenge in the labor market: toward new opportunities

Labor markets were always studied on the basis of statistical data – with indicators describing the employment phenomena – and administrative data, providing information on the trends and dynamics of the market.

The main issue for policymakers who deal with labor markets is the matching between demand and supply of labor. While the labor **supply** is sufficiently described, and we have data of the different nature of unemployment, profiles, and skills of jobseekers, we do not have enough data on the labor **demand**, i.e., what companies look for, both the number of vacancies the skill requirements.

Knowledge of the labor demand is crucial, especially for the following characteristics of data:

- Timeliness
- Territorial granularity
- Links between occupations and related skills

Until now, surveys were the main source of information on labor demand. However, relying on surveys had some drawbacks:

- They are expensive to carry out;
- They are difficult to carry out, therefore they cannot have a high frequency, and often do not offer information on the local dimension or suffer from problems of under-sampling;
- The final data that results could be already obsolete due to the time-lag between the start of the distribution and the end of the analysis;
- It has a top-down approach, i.e., soft skills and professional expertise are usually pre-defined.

From the stakeholders' perspective, there is a need for environmental knowledge, i.e., data to describe the framework and to supply context information. The stakeholders need to realize that data-driven systems are able to provide more precise and timely information and to support the decision-making process in an effective way. ARLI, a project carried out for the European Commission[2], underlined the information needs of the regional and local stakeholders who were interviewed during the project: information on the current and future developments in the labor market is needed for analysis on regional, local, and district levels and/or for analyzing specific sectors and professional groups. The requested data especially concern the labor demand, and the effective matching between demand and supply (both quantitative and qualitative).

There is another place where the demand for labor is expressed; the Web is becoming an increasingly important channel for posting jobs and, in general, for the matching between demand and supply of labor. There is a multitude of portals (a.k.a. **Job Boards**) that publish Job Vacancies.

We can identify three main advantages that Web Job Vacancies provide with respect to survey-based analyses:

- **Data Scraping:** the cost of collecting data (*aka* data gathering or scraping) is lower (specifically, moderate initial costs, but low marginal costs).
- **Time to market:** data are always up-to-date regarding the observed phenomenon, and this enables the use of real-time analysis techniques.
- **Bottom-up approach:** the classification is richer as it emerges from the data rather than from a pre-defined model or taxonomy. This peculiarity is particularly useful for identifying soft and professional skills, especially if we consider that the skill mismatch has a lot to do with the difficulty of defining and classifying skills properly. Unlike traditional surveys, the skills expressed online do not have to fit pre-defined classifications, like in a questionnaire, but they are expressed freely (i.e., they are not strategic responses to sensitive questions).

2 Achieving Regional and Local Impact (ARLI) funded from the European Union's Progress Programme, 2013-2014 (www.regionallabourmarketmonitoring.net/arli_public.htm).

It is worth noting that Web data sources have an unexpressed informative power that allows one to gather a lot of information about both labor demand and supply. Consequently, this process involves several challenging tasks that we will briefly discuss in this paper, such as data selection, data transformation and cleaning, data reasoning and mining, as well as data visualization.

An interesting aspect to monitor is the introduction of a new paradigm: from "pre-defined answers" (precise data, structured, collected ad-hoc, and small) to "let the data speak" – the "data-driven economy or technology" (huge amount of unstructured data, presence of inaccuracy and analysis scalability). To let the data speak means to encourage an approach to the analysis of the phenomena based on the questions that arise from the observation of correlations between objects or "hidden" facts, hard to know *a priori* and without the "data-driven" observation of the reality of interest. Of course, another huge difference – not negligible – in this kind of Big Data approach is the real-time collection of all (or almost all) data.

The representativeness of web data

A key challenge in using online job vacancies is ascertaining whether the set of online job vacancies is a representative sample of all job vacancies in a specified economy. Even if it should be considered finite, the population of job vacancies at any given moment in time is not easily counted nor is its structure easy to determine. Moreover, job vacancies are voluntary, jobs and people are reallocated in the labor market and in firms, tasks are split, restructured, or partially switched and recruitment strategies might have sectoral and occupational specificities (Mang 2012; De Pedraza et al. 2007; Gosling et al. 2004).

Statistically speaking all these elements would suggest to consider online vacancy data as containing missing not at random (NMAR) observations (Little and Rubin 1987); thus, available samples are prone to problems of self-selection and/or non-response. Various approaches have been taken by different authors in attempting to deal with representativeness issues in the setting of online job vacancy. Some researchers have used information from the supply side of the market and assessed the coverage of online vacancies based on the sectoral and occupational structure of LFS data (Štefánik

2012; Jackson 2007; Steinmetz et al. 2009). Others focused on segments of the labor market characterized by widespread access to the Internet (e.g., ICT) where coverage bias is likely to be a minor problem (Kennan et al. 2006; Wade and Parent 2001).

Kureková et al. (2016) consider online data generalizable due to the dominant market share and very high reputation of the chosen portal among employers and employees. Others suggest that, depending on the particular research focus, online job vacancies could be coupled with other sources of vacancy data or text describing analyzed professions. Wade and Parent (2001) acknowledge that coverage bias can be addressed by complementing their methodology with structured interviews with employers or recruiters, whereas Kureková et al. (2016) suggest using information from the EURES website because of its standardized platform and relatively wide usage across European countries. Overall weighting as an adjustment technique (Steinmetz et al., 2009; De Pedraza et al. 2007; Eurostat 2010) can be implemented in improving the representativeness of online job vacancy data.

More specifically, statistical methods anchored in the literature on missing data and self-selection within a model-based (Valliant et al. 2000) approach (such as Post-stratification weighting, Propensity score adjustment and Endogenous/Heckman selection models) to the correction of data not missing at random, could be used to adjust biases stemming from the structure of online vacancy data. It is important to note that online data can be usefully employed in official statistics even if they are not entirely representative. In fact, their evolution over time can mimic quite closely that of the entire population, and be used to construct a dynamic index.

The two major examples in this respect are from the US. The Billion Price Project at MIT monitors billions of online prices around the world. In the US, although online prices are not representative of the basket of goods used in the CPI, their change over time predicts CPI inflation extremely well. For this reason, the Bureau of Labor Statistics is collaborating with MIT to introduce online data checking in inflation computations. Another relevant example is from the labor market. Vacancy data are crucial indicators for labor market slack. In the US two main data sources are used in official statistics and for policymaking: the Job Openings and Labor Turnover Survey

(JOLTS), administered by the Bureau of Labor Statistics, and the Conference Board Help Wanted Online series (HWOL) which monitors online vacancies (Barnichon 2010). Although not comparable in terms of representativeness (the first is designed to be representative, the second is not) both series have a remarkably close evolution over time and are jointly used in policymaking. However, the debate on the complete reliability of online vacancies is still very hot: for example, one of the reasons why the FED postponed the interest rate increase in June 2016 was the weak conditions in the labor market described by the HWOL series whereas Cajner and Ratner (2016) show that there was a bias in online data which led to spurious results.

How to deal with web job vacancies

As the job vacancies posted on the Web are expressed through semi-structured or unstructured texts, they require rigorous work from a scientific, methodological, and technical point of view if we want to extract knowledge (and, subsequently, value-added) from these data (Mezzanzanica 2013). A process that describes all the steps required for dealing with Web Job Vacancies is the Knowledge Discovery in Databases (KDD), defined by Fayyad et al. (1996) as the "Process for Extracting Useful Knowledge from Volumes of Data". They explicate: "*The value of storing volumes of data depends on our ability to extract useful reports, spot interesting events and trends, support decisions and policy based on statistical analysis and inference, and exploit the data to achieve business, operational, or scientific goals.*" *This process clearly relies on "[...] our ability to extract useful reports, events and trends, support decisions and policy based on statistical analysis and inference.*" (Fayyad et al. 1996).

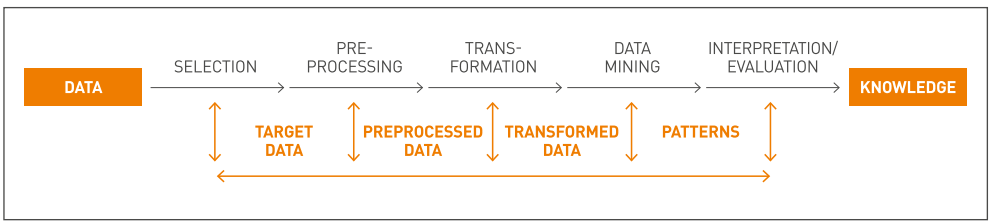

Figure 3: The KDD process *(Source: Fayyad et al. (1996).)*

Figure 3 summarizes the process of data extraction and its different components, underlining the complexity of data quality, data significance, selection of sources for creating a digital observatory, and, consequently, the issues of web scraping and setting up a data model for classifying the information. We can summarize these steps as follows:

1. **Data Selection:** Data are scraped from a pre-defined list of Web sources. That requires dealing with data significance and source reliability issues. Furthermore, a multi-criteria decision-making approach should be used for identifying and ranking data sources on the basis of their main characteristics.
2. **Data transformation and cleaning:** This is a mandatory task due to the heterogeneity of the data, characterized by different (unknown) data models with poor data quality. Indeed, data quality is usually defined as "fitness for use" and includes many dimensions (e.g., consistency, completeness, soundness) at both data and structure level. We addressed and developed a methodology for assessing and cleaning structured data in a formal and automatic way (Mezzanzanica et al. 2015, Boselli et al. 2014). However, this task becomes challenging in the case of semi-structured/unstructured data, mainly due to the lack of a rigorous data model, which prevents the use of the well-known ETL techniques.
3. **Data reasoning and mining**: This task aims to extract the value from the data in answering a specific business or research question (see Amato et al. 2015 for details).
4. **Data visualization**: This generally refers to the representation of data/knowledge or the data evolution over time (a.k.a. dynamics) by means of infographics, having in mind the stakeholder peculiarities and abilities to reading the data. Nowadays, the definition and use of narrative paradigms that support different users in understating the data has become a challenge.

Conclusion

Web data have received much attention from both industrial and academic communities. This results from their informative power enabling the description of complex phenomena that evolve dynamically and continuously over time, as in the case of web job vacancies. These data are frequently heterogeneous and involve a mixture of semi-structured and unstructured data with different sizes and degrees

of granularity. All these data characteristics play a crucial role in any "Big Data" application.

Indeed, a real-time analysis of web job vacancies collected by heterogeneous and unstructured data sources makes it possible to obtain fine-grained and up-to-date information about labor demand trends, identifying the skills requested by the market, and supporting the decision-making activities and strategies of labor market operators. As one might imagine, this process is far from straightforward as it involves statisticians, economists, and computer scientists working in close cooperation with application domain experts, each focusing on their own perspective.

Our research progress in this direction, together with our expertise in the labor market domain, suggests that the use of web job vacancies can give a competitive advantage to labor market observatories in understanding, monitoring, and explaining labor market phenomena, thus supporting the decision-making activities of labor market operators in a more effective and timely way.

Silvia Dusi is a researcher at Interuniversity Research Centre on Public Services (CRISP) in Milan, where she has worked since 2010. She holds a Master of Science in Economics from the University of Milano Bicocca. To support CRISP activities, she deals with the center's business development, following the international projects, being the contact person for the European agencies, and enlarging the network of European partners.

The new world of work is characterized by globalized employment, a mobile yet vulnerable workforce, and the challenges of demography and rising income inequality. Technological changes in both the demand for and supply of skills have a cross-cutting influence on how labor markets develop. In this book, different stakeholders from international organizations in the private and public sector discuss which role Public Employment Services and Workforce Development Agencies ought to play in the labor market today and in the future, why cooperation is crucial, and what kind of support digital services and software can provide for a more effective and efficient delivery.

Managing Workforce Potential – A 20/20 Vision on the Future of Employment Services seeks to inspire decision-makers in and around the labor market to reflect on governance, services, and partnerships to better cater to the new world of work.

Why this book?

As a world leader in Public Employment software solutions, WCC believes in sharing knowledge. It is our vision that combining what we know and sharing this with the world leads to maximum value across the board. This is why we take initiatives to both exchange and expand expertise. For example, we started the PEPTalk webinar series, which provides a platform for Public Employment Services to share their knowledge about best practices and their vision on the labor market. This book is another example; with its publication, we aim to contribute to an all-round clearer vision on the developments in public employment.

*The term **20/20 vision** is used to express normal sharpness of vision. It means you can see clearly at 20 feet what should normally be seen at that distance.*

ISBN 978 90 8252 531 1