TOWARDS SMART STATISTICS IN LABOUR MARKET DOMAIN

Inna Novalija
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia

inna.koval@ijs.si

Marko Grobelnik
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia

marko.grobelnik@ijs.si

ABSTRACT

In this paper, we present a proposal for developing smart labour market statistics based on streams of enriched textual data and illustrate its application on job vacancies from European countries. We define smart statistics scenarios including demand analysis scenario, skills ontology development scenario and skills ontology evolution scenario. We identify stakeholders – consumers for smart statistics and define the initial set of smart labour market statistical indicators.

KEYWORDS

Smart statistics, labour market, demand analysis.

1. INTRODUCTION

An essential feature of modern economy is the appearance of new skills, such as digital skills. For instance, e-skills lead to the exponential increases in production and consumption of data.

While job profiles vary and are still in the process of being defined, organizations agree that they need the new breed of workers.

Accordingly, the European institutions take major initiatives related to digitalization of labor market, training of new skills and meeting the labour demand.

Historically, the labour market statisticians use standard measures of the labour demand and labour supply based on traditional surveys – job vacancy surveys, wage survey, labour force surveys. The unemployment rate provides information on the supply of persons looking for work in excess of those who are currently employed. Data on employment provide information on the demand for workers that is already met by employers.

The data-driven smart labour market statistics intends to:

- use the available historical job vacancies data,
- use the available real-time job vacancies data,
- use the available real-time and historical dataset of additional data (described below),
- align data sources,
- construct models and obtain novel smart labour market indicators that will complement existing labour market statistics,
- provide a system for delivering results to the users.

The smart labour market statistics approach will combine advanced data processing, modelling and visualization methods in order to develop trusted techniques for job vacancies analysis with respect to defined scenarios – demand analysis, skills ontology development and skills ontology evolution.

2. BACKGROUND

The development of smart labour market statistics touches a number of issues from labour market policies area and would provide contributions to questions related to:

- job creation,
- education and training systems,
- labour market segmentation,
- improving skill supply and productivity.

For instance, the analysis of the available job vacancies could offer an insight into what skills are required in the particular area. Effective trainings based on skills demand could be organized and that would lead into better labour market integration.

A number of stakeholder types will benefit from the development of smart labour market statistics. In particular, the targeted stakeholders are:

- Statisticians from National and European statistical offices who are interested in the application of new technologies for production of the official statistics.
- Individual persons who are searching for new employment opportunities. In particular, individuals are interested in the job vacancies that are compatible with their current skills and in the methods (like trainings) providing the possibilities to obtain new skills in demand.
- Public and private employment agencies interested in up-to-date employees profiles.
- Education and training institutions from different levels and forms of education - general/vocational education, higher education, public/private, initial/ adult education. Educational institutions are interested in relevant skills and topics that should be part of the curriculum programs.
- Ministries of labour/manpower, economy/industry/trade, education, finance, etc. The policy makers, such as ministries, are interested in the overall labour market situation, with respect to location and time, in the labour market segmentation and in the processes of improving supply and productivity.
- Standards development organizations. National or International organizations whose primary activities are developing, coordinating, promulgating, revising, amending, reissuing, interpreting, or otherwise producing technical standards that are intended to address the needs of some

relatively wide base of affected adopters. Interested in new technologies developed in relation to labour market.

- Academic and research institutes. Public and private entities who conduct research in relevant areas. Research institutions are interested in the development of novel methodologies and usage of appearing new data sources.

3. RELATED WORK

The European Data Science Academy (EDSA) [1] was an H2020 EU project that ran between February 2015 and January 2018. The objective of the EDSA project was to deliver the learning tools that are crucially needed to close the skill gap in Data Science in the EU. The EDSA project has developed a virtuous learning production cycle for Data Science, and has:

- Analyzed the sector specific skillsets for data analysts across Europe with results reflected at EDSA demand and supply dashboard;
- Developed modular and adaptable curricula to meet these data science needs; and
- Delivered training supported by multiplatform resources, introducing Learning pathway mechanism that enables effective online training.

EDSA project established a pipeline for job vacancy collecting and analysis that will be reused for the purpose of smart statistics.

An ontology called SARO (Skills and Recruitment Ontology) [2] has been developed to capture important terms and relationships to facilitate the skills analysis. SARO ontology concepts included relevant classes to job vacancy datasets, such as Skill and JobPosting. Examples of instances of class Skill would be skills, such as “Data analysis”, “Java programming language” et al.

ESCO [3] is the multilingual classification of European Skills, Competences, Qualifications and Occupations. It identifies and categorizes skills/competences, qualifications and occupations relevant for the EU labour market and education and training, in 25 European languages. The system provides occupational profiles showing the relationships between occupations, skills/competences and qualifications. For instance, one example of existing ESCO skill is “JavaScript” (with alternative labels “Client-side JavaScript”, "JavaScript 1.7" et al.).

Both SARO and ESCO ontologies are useful for the aim of smart statistics, in particular for skills ontology development and skills ontology evolution scenarios. However, the ontologies usually are manually manipulated, and the methods developed for smart labour market statistics should overcome the difficulties related to this issue. The ontology evolution scenario of smart labour market statistics envisions automatic identification of emerging and decreasing skills from the data perspective.

4. PROBLEM DEFINITION

4.1 DATA SOURCES

The main data sources available for the development of smart labour market statistics are historical and current data about job vacancies in the area of digital technologies and data science around Europe (~5.000.000 job vacancies 2015-2018).

Additional data sources may include:

- Social media data, such as news, Twitter data that might be relevant for labour market.

- Labour supply data (based on user profile analysis).

Open job vacancies can be found using job search services. These services aggregate job vacancies by location, sector, applicant qualifications and skill set or type. One such service is Adzuna [4], a search engine for job ads, which mostly covers English-speaking countries.

For data acquisition and enrichment, dedicated APIs, including Adzuna API, are used, as well as custom web crawlers are developed. The data is formatted to JSON to aid further processing and enrichment. The job vacancy dataset is obtained with respect to trust and privacy regulations, the personal data is not collected.

Job vacancies usually contain the information, such as job position title, job description, company and job location. In such way, job vacations that are constantly crawled/web-scraped present a data stream. The job title and job description are textual data that contain information about skills that employee should have.

On the obtained data wikification - identifying and linking textual components (including skills) to the corresponding Wikipedia pages [5] is performed. This is done using Wikifier [6], which also supports cross and multi-linguality enabling extraction and annotation of relevant information from job vacancies in different languages. The data is tagged with concepts from GeoNames ontology [7]. To job postings where latitude and longitude have been available, GeoNames location uri and location name are added. To the postings where only location name has been available, the coordinates and location uri are added.

The job vacancy data representation level depends on the specific country. For the United Kingdom, France, Germany and the Netherlands there is a substantial collection of job vacancies in the area of digital technologies.

4.2 CONCEPTUAL ARCHITECTURE

The labour market statistics conceptual structure is built upon the following major blocks:

1. Data sources related to different aspects of smart labour market. The main data source aggregates historical and current job vacancies in the area of digital technologies and data science around Europe.

2. Modelling smart labour market statistics takes central part of the smart labour market statistics approach, where the goal is to construct models based on different data sources, updated in business-real-time (as needed or as data sources allow). Models shall bring understanding of the smart labour market statistics domain and shall be used for aggregation, ontology development and ontology evolution.

3. Targeted users are smart statistics consumers. There are several major groups of users (described above). The example users might include statisticians, policy makers, individual users (residents and non-residents), training and educational organizations and other.

4. Finally, applications of smart labour market statistics are multiscale - they can be presented at cross-country level (around

Europe) country level (UK, France, the Netherlands etc.), city/area level and conceptual level (ontology).

Figure 1 illustrates the conceptual architecture diagram for smart labour market statistics.

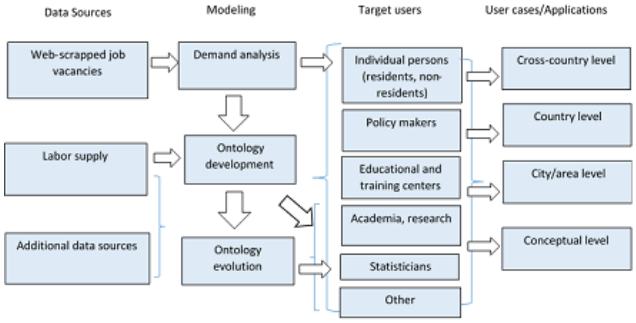


Figure 1: Conceptual Architecture

The key characteristics of the development techniques will include:

- Interpretability and transparency of the models – the aim is, for a model to be able to explain its decision in a human readable manner (vs. black box models, which provide results without explanation).

- Non-stationary modelling techniques are required due to changing data and its statistical properties in time. For instance, the ontology evolution process will be modeled taking to the account the incremental data arriving to the system.

- Multi-resolution nature of the models, having the property to observe the structure of a model on multiple levels of granularity, depending on the application needs.

- Scalability for building models is required due to the nature of incoming data streams.

4.3 SCENARIOS

The smart labour market statistics proposal includes three scenarios - demand analysis scenario, ontology development scenario and ontology evolution scenario described below.

4.3.1 DEMAND ANALYSIS

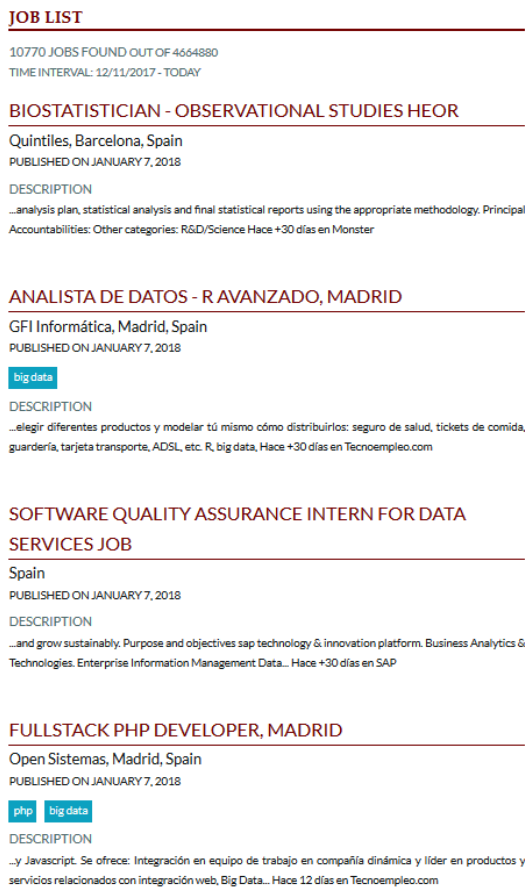
Demand analysis scenario suggests production of statistical indicators based on the available job vacancies using techniques for data preprocessing, semantic annotation, cross-linguality, location identification and aggregation.

Job vacancies in structural and semi-structural form are the input to into the system, while statistics related to overall job demand, job demand with respect to particular location, job demand with respect to particular skill (skill demand) and time frame are the outputs of the system.

Figure 2 presents an example of crawled and processed job vacancies.

4.3.2 SKILLS ONTOLOGY DEVELOPMENT

Ontologies reduce the amount of information overload in the working process by encoding the structure of a specific domain and offering easier access to the information for the users. Gruber [8] states that an ontology defines (specifies) the concepts, relationships, and other distinctions that are relevant for modeling a domain. The specification takes the form of the definitions of representational vocabulary (classes, relations, and so forth), which provide meanings for the vocabulary and formal constraints on its coherent use.


JOB LIST

10770 JOBS FOUND OUT OF 4664880
TIME INTERVAL: 12/11/2017 - TODAY

BIOSTATISTICIAN - OBSERVATIONAL STUDIES HEOR

Quintiles, Barcelona, Spain
PUBLISHED ON JANUARY 7, 2018

DESCRIPTION
...analysis plan, statistical analysis and final statistical reports using the appropriate methodology. Principal Accountabilities: Other categories: R&D/Science Hace +30 días en Monster

ANALISTA DE DATOS - R AVANZADO, MADRID

GFI Informática, Madrid, Spain
PUBLISHED ON JANUARY 7, 2018

big data

DESCRIPTION
...elegir diferentes productos y modelar tú mismo cómo distribuirlos: seguro de salud, tickets de comida, guardería, tarjeta transporte, ADSL, etc. R, big data, Hace +30 días en Tecnoempleo.com

SOFTWARE QUALITY ASSURANCE INTERN FOR DATA SERVICES JOB

Spain
PUBLISHED ON JANUARY 7, 2018

DESCRIPTION
...and grow sustainably. Purpose and objectives sap technology & innovation platform. Business Analytics & Technologies. Enterprise Information Management Data... Hace +30 días en SAP

FULLSTACK PHP DEVELOPER, MADRID

Open Sistemas, Madrid, Spain
PUBLISHED ON JANUARY 7, 2018

php big data

DESCRIPTION
...y Javascript. Se ofrece: Integración en equipo de trabajo en compañía dinámica y líder en productos y servicios relacionados con integración web, Big Data... Hace 12 días en Tecnoempleo.com

Figure 2: Example of Job Vacancies Crawled and Processed

Ontologies are often manually developed and maintained, what requires a sufficient user efforts.

In the ontology development scenario an automatic (or semi-automatic) bottom-up process of creating ontology from available job vacancies will be suggested.

The relevant skills (extracted from the job vacancies) will be defined and formalized. Using semantic annotation and cross-linguality techniques for skills extraction based on JSI Wikifier tool [6] will enable the possibility of including the newest available skills “on the market” that are not yet captured in the ontologies, taxonomies and classifications that are manually developed. The input to the ontology development scenario is a set of job vacancies and the output is ontology of skills presenting the domain structure that can be compared to or used for official classifications.

4.3.3 SKILLS ONTOLOGY EVOLUTION

Ontology Evolution is the timely adaptation of an ontology to the arisen changes and the consistent propagation of these

changes to dependent artefacts [9]. Ontology evolution is a process that combines a set of technical and managerial activities and ensures that the ontology continues to meet organizational objectives and users' needs in an efficient and effective way.

Ontology management is the whole set of methods and techniques that is necessary to efficiently use multiple variants of ontologies from possibly different sources for different tasks [10].

Scenario 3 will suggest an automatic (or semi-automatic) ontology evolution process based on the real-time job vacancy stream. With respect to the nature of job vacancy data stream and skills extracted from job it will be possible to see the dynamics of evolving skills – when the new skills (not included into the current ontology versions appear) and how the skills ontology is changing with time.

In particular, it could be possible to observe appearing new skills and suggest them for inclusion into official skills classifications. In addition, it could be visible how fast the ontology changes, which could be the indicator of the technological progress on the relevant market.

For instance, the current version of ESCO classification does not contain "TensorFlow" skill (TensorFlow [11] is an open-source software library for dataflow programming across a range of tasks, appeared in 2015). TensorFlow, which is already present in job vacancies, could be captured during ontology evolution process and suggested as a new concept for official classifications.

5. STATISTICAL INDICATORS

Traditionally the indicators related to labour market have been based on survey responses. The smart labour market statistics proposal introduces a possibility to complement standard statistical indicators, such as job vacancy rate with novel "data inspired" knowledge.

The smart labour market statistics indicators use data sources, previously not covered by official statistics, and in such way complementary to traditional data sources. The smart labour market statistics indicators are based on real-time data streams, which makes possible to obtain not only historical, but also current values for job vacancies that could be used for different purposes, such as nowcasting. In addition, the smart labour market statistics indicators take into the account data cross-lingual and multi-lingual nature of streaming data and can be produced at the multiscale levels – cross-country, country, city (area) levels.

The scenarios described above would result into a number of smart labour market indicators with multiscale options. In particular:

- Up-to date job vacancies statistics on a cross-country/country/city(area) level. Example: job vacancies in UK and France in the last month
- Up-to date skills statistics on a cross-country/country/city(area) level. Example: top 10 skills in UK in the last month
- Up-to date location statistics. Example: top locations for specific skill
- Ontology development statistics. Example: number of concepts in the ontology
- Ontology evolution statistics. Example: emerging skills in the ontology in the last 3 months

Since the data has a streaming nature, different kinds of multiscale and aggregation options can be handled with respect to time parameters.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented a proposal for developing smart labour market statistics based on streams of enriched textual data, such as job vacancies from European countries. We define smart statistics scenarios, such as demand analysis scenario, skills ontology development scenario and skills ontology evolution scenario. The future work would include the implementation of the smart labour market scenarios, quality assessment and evaluation of the produced statistical outcomes.

7. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and EDSA European Union Horizon 2020 project under grant agreement No 64393.

8. REFERENCES

[1] EDSA, http://edsa-project.eu (accessed in August, 2018).

[2] Sibarani, Elisa & Scerri, Simon & Mousavi, Najmeh & Auer, Sören. (2016). Ontology-based Skills Demand and Trend Analysis. 10.13140/RG.2.1.3452.8249.

[3] ESCO taxonomy, https://ec.europa.eu/esco/portal (accessed in August 2018).

[4] Adzuna developer page, https://developer.adzuna.com/overview (accessed in August, 2018).

[5] Ratinov, L., Roth, D., Downey, D. and Anderson, M. Local and global algorithms for disambiguation to wikipedia. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 1375–1384. Association for Computational Linguistics, 2011.

[6] JSI Wikifier, http://wikifier.org (accessed in May, 2018).

[7] GeoNames ontology, http://www.geonames.org/ontology/documentation.html (accessed in August, 2018).

[8] Ontology (by Tom Gruber), http://tomgruber.org/writing/ontology-definition-2007.htm (accessed in August, 2018).

[9] M. Klein and D. Fensel, Ontology versioning for the Semantic Web, Proc. International Semantic Web Working Symposium (SWWS), USA, 2001

[10] L. Stojanovic, B. Motik, Ontology evolution with ontology, in: EKAW02 Workshop on Evaluation of Ontology-based Tools (EON2002), CEUR Workshop Proceedings, Sigüenza, vol. 62, 2002, pp. 53–62

[11] TensorFlow, https://en.wikipedia.org/wiki/TensorFlow (accessed in August, 2018).