UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'ENERGIA DEI SISTEMI,
DEL TERRITORIO E DELLE COSTRUZIONI

RELAZIONE PER IL CONSEGUIMENTO DELLA
LAUREA MAGISTRALE IN INGEGNERIA GESTIONALE

A Novel Approach to Define the Impact of Technological Changes on Skills and Job Profiles

SINTESI

RELATORE

Prof. Ing. Gualtiero Fantoni
Dipartimento di Ingegneria Civile e Industriale,
Università di Pisa

CORRELATORE

Dott. Ing. Filippo Chiarello
Dipartimento di Ingegneria dell'Energia. dei Sistemi,
del Territorio e delle Costruzioni,
Università di Pisa

IL CANDIDATO

Pietro Manfredi

p.manfredi95@gmail.com

Sessione di Laurea del 27/11/2019
Anno Accademico 2018/2019
Consultazione NON consentita

A Novel Approach to Define the Impact of Technological Changes on Skills and Job Profiles

Pietro Manfredi

Sommario

La presente tesi ha l’obiettivo di definire una metodologia semi-automatica al fine di restituire dati continuamente aggiornati sul mercato del lavoro e predirne i futuri cambiamenti. L’approccio proposto parte dall’analisi del progresso tecnologico all'interno di un determinato campo o settore di interesse, utilizzando articoli scientifici come fonte primaria di dati. Questa prima fase ha un duplice output: un grafo di tecnologie (la cui grandezza dei nodi è proporzionale alla importanza delle stesse, mentre i rami ne esplicitano le connessioni) e una visualizzazione della crescita (o decrescita) di interesse delle stesse. Successivamente, attraverso la ricerca delle tecnologie suddette all’interno di corsi online (MOOC) e sistemi di classificazione delle professioni (ESCO e O*NET), viene identificato il legame con i rispettivi profili professionali e competenze. Infine, viene fornita una misura quantitativa della rilevanza strategica dei profili professionali identificati, la cui interpretazione rende questa metodologia un potenziale strumento per la gestione ottimizzata delle risorse umane.

Abstract

This thesis focuses on the development of a new methodology, driven by automated techniques and able to provide real-time data of the labour market and predict its future changes. The proposed approach starts from the analysis of the technological progress within a given field or sector of interest, using scientific articles as a primary source of data. This first phase has a dual output: a network diagram, which provides an understanding of how technologies are connected to each other as well as their importance, and a representation of their growth of interest over time. Subsequently, technologies are linked to skills and job profiles through massive open online courses (MOOCs) in the first case, and occupational frameworks (e.g. ESCO and O*NET) in the second. In the end, this thesis develops a way of defining a quantitative measure of importance for each occupation whose interpretation makes this methodology a potential tool for the strategic management of human resources.

1 Introduction

There is a need today to detect emerging technologies and related skills in order to align the work demand to the existing competencies.

The problem of identifying future skills is becoming increasingly challenging given the current dynamics present in the global economy. Being able to understand job requirements and react fast to changes in skill needs is the key to success on the labour market[1]. In addition, the comprehension of changes taking place is not only crucial for companies but also citizens, employers, training providers and policy makers since a more detailed knowledge of skills requirements helps in designing training programmes giving the opportunity to upskill and reskill.

2 State of the art

Skill anticipation has been carried out for decades now, however with the advent of machine learning, big data analysis and artificial intelligence algorithms, it is possible to extract the needed information in a timely manner. Automated techniques are in part already being used to anticipate changes in the labour market, but with differences in terms of skill definition, time span, data source, adopted methodology and the national, regional or sectoral scope. However, from a methodological point of view, these approaches can be classified into two groups:

- *Group A*: Exclusive use of *available online data* to which automated techniques are applied to interpret labour market developments;
- *Group B*: Information firstly gained through surveys and expert consultations (human based methodologies for data gathering) to which automated techniques are applied to anticipate changes in skills and occupations;

Burning Glass[2] (leader in job matching and labor market analytics solutions), *Upwork*[3] (global freelancing platform) and *OVATE*[4] (Online Vacancy Analysis Tool for Europe) are major examples of *group A* where skill anticipation is done through the analysis of online job

[1] Dench, S. (1997). Changing skill needs: what makes people employable?. *Industrial and commercial training*, *29*(6), 190-193.

[2] https://www.burning-glass.com/research/

[3] https://www.upwork.com/press/2019/05/14/q1-2019-skills/

[4] European Centre for the Development of Vocational Training (Cedefop). (2019). Online job vacancies and skills analysis: a Cedefop pan-European approach

postings. More precisely, data on most requested skills and occupations is obtained by analysing how many job offers are present for each job profile and how many times a certain skill is mentioned. Approaches belonging to *group B*, are more focused on skill foresight rather than providing insight into the current state of the labour market. Examples are Frey and Osborne’s work and the study carried out by Pearson and Nesta (The future of skills: employment 2030). In both cases, experts give their opinion on possible future scenarios of existing occupations, and subsequently a machine learning model is used to anticipate their change over the next decade or so.

Even though these approaches from group A and B are effective and provide valuable information for stakeholders, attention still needs to be paid when using their output to anticipate future scenarios as:

- Bias in results can be present in both groups, not only in experts’ judgment but also in online job postings as job-specific skills may be taken for granted by employers and the emphasis put on transversal competencies instead;
- Job vacancies reflect today’s scenario leaving future interpretation to the stakeholder;
- All analysed approaches mainly rely on a single data source;

3 Methodology

Following these considerations, I have developed a new methodology which is objective, future proof and not reliant on a single data source. For it to be future oriented, focus must be first placed on key drivers of change which impact the labour market. The latter are many, but the most significant involve technological change. New technologies require new or updated current skills thus shaping today's occupations or creating new ones. This problem can be better approached by defining specific questions for which we need an answer.

1. How are *existing technologies* changing? Are there *emerging technologies*?
2. How is *technological evolution* impacting *skills* and *job profiles*?
3. Which *skills* and *job profiles* have and will have a higher *request* in the near future?

From a methodological point of view, however, these questions can only be answered exhaustively once a process to extract *technologies*, *skills* and *job profiles* has been found. In detail, once growing technologies have been detected, it is then possible to state that related job profiles and skills will have a higher demand in the future.

The workflow used is shown in *figure 1.*

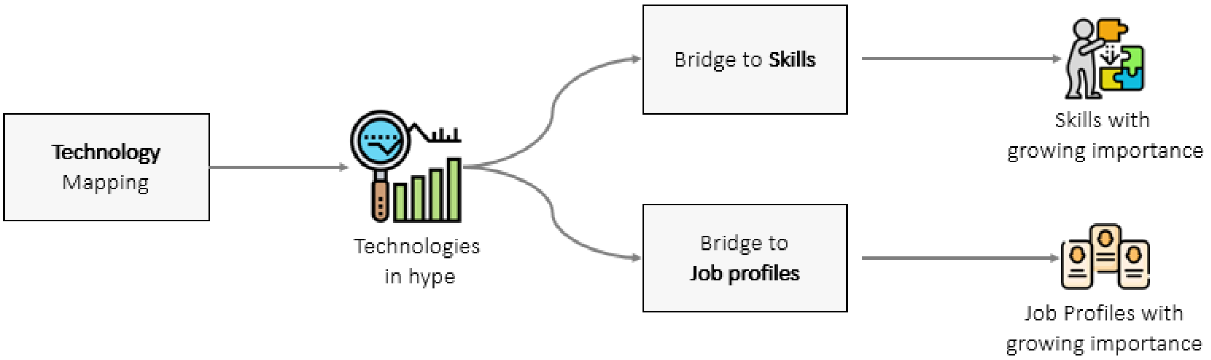


Figure 1 - Workflow used to estimate future changes in Labour market

4 Technology mapping

There are various data sources from which technological information can be extracted, for instance Wikipedia[5], papers and patents. The approach I have focused on, uses published papers (technical papers and non-technical) as the main source of information. Technologies can be extracted in two ways: by retrieving all keywords linked to papers (author's keywords and keywords chosen by the journal on which the article is published) and secondly manually filtering those keywords which are not inherent by any means to technologies, or by simply using a pre-existing list of technologies and extracting those that are found in the papers' abstracts. Technologies are then visualized through a network diagram[6] (using VOSviewer[7]). To show a practical example, I have applied the described approach to papers related to Data science in the robotics field. These articles have been retrieved from Web of Science (similar to Scopus) which gives the possibility to bulk download papers, using a research query[8] containing synonyms and terms inherent to the field of interest (data science in the robotics field).

Figure 2 displays the output of this process.

[5] Fantoni, G., Chiarello, F., Fareri, S., Pira, S., & Guadagni, A. (2018). Defining industry 4.0 professional archetypes: a data-driven approach. *Economy, employment and skills: European, regional and global perspectives in an age of uncertainty*, 75.

[6] A network diagram is a form of data visualization made of nodes, which resemble specific entities, and arcs which link the nodes.

[7] https://www.vosviewer.com/

[8] A query is a set of terms or strings that are linked together through the use of logical connectors (e.g. AND or OR) that allow a targeted search on specific databases.

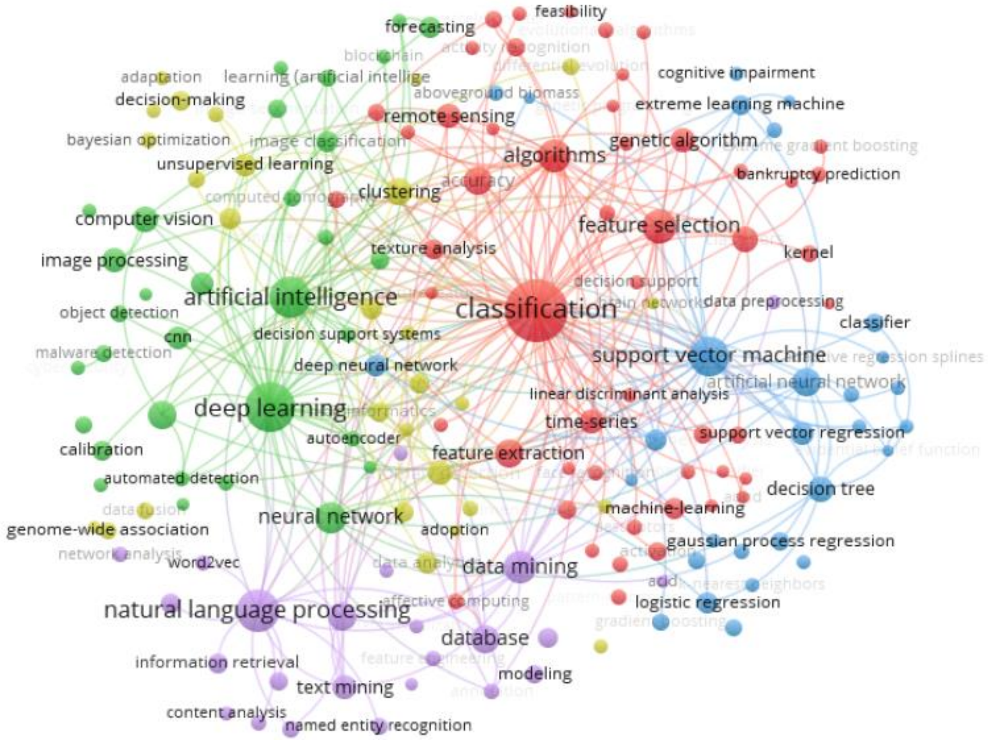


Figure 2 - Network of technologies present in the robotics field

Network diagrams are not only a "cool" way of visualizing data, but in this specific case they can be used to highlight how technologies are connected to each other and how relevant they are. In detail, the importance of technologies and related interconnections are respectively identified by the size of each node (the greater the size, the greater the importance of a technology) and arcs. An arc exists between two technologies if they appear in the same paper. For instance, in *figure 2,* technologies with high relevance are "classification systems", "deep learning", "natural language processing".

However, network diagrams provide a static view of technologies, thus not showing their evolution, and for this reason, a trend analysis is carried out by counting the number of times each technology appears in papers over time and normalized within the domain (temporal data is obtained by using the publication year of the articles). The output, which is stored in an excel file, is then imported in R studio[9] and visualized using the ggplot[10] package.

[9] https://rstudio.com/

[10] ggplot2 is a data visualization package for the statistical programming language R.

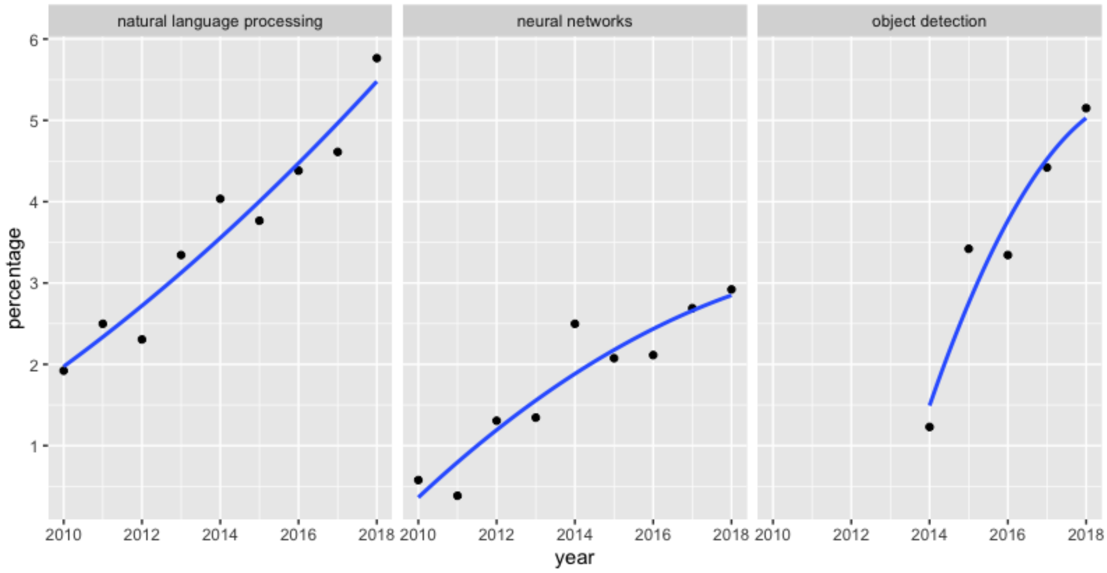


Figure 3 - Trend analysis of technologies related to the Robotics field (the y-axis represents the percentage of papers related to a specific technology)

Trends provide a better understanding of technological evolution, for instance technologies which might not appear as relevant in the network diagram (nodes which are small in size) could still be important as their trends could be growing in time. For instance, *figure 3* shows that object detection, which in *figure 2* is a small node, thus apparently not relevant, is not only considered as an emerging technology since it only appears in papers after the year 2014, but also growing in importance. That said, I suggest using network diagrams as well as trends for a complete analysis on technologies and related changes.

5 Bridging to skills

After identifying technologies, the next step regards the extraction of skills. As previously mentioned, there is a direct connection between the two concepts, a growing technology translates into high request of related competencies. Knowledge of the latter can be found in CVs, job descriptions (even though they are not easily accessible), online job offers, MOOCs (massive open online courses), papers (by searching for phrases mentioning a skill) and occupational frameworks[11] (e.g. ESCO[12] or O*NET[13]). The best and most effective way I suggest to directly link technologies to skills, is by using MOOCs since competencies are mostly gained through a learning process. In addition, every course has a well-defined structure, generally called *syllabus* making the data processing easier and less time

[11] Databases which contain detailed information on competencies and job occupations.

[12] European classification of job profiles and competencies (https://ec.europa.eu/esco/portal/home).

[13] US classification of job profiles and competencies (https://www.onetonline.org/).

consuming. In detail the workflow developed to process MOOC data for the extraction of skills is shown in *figure 4.*

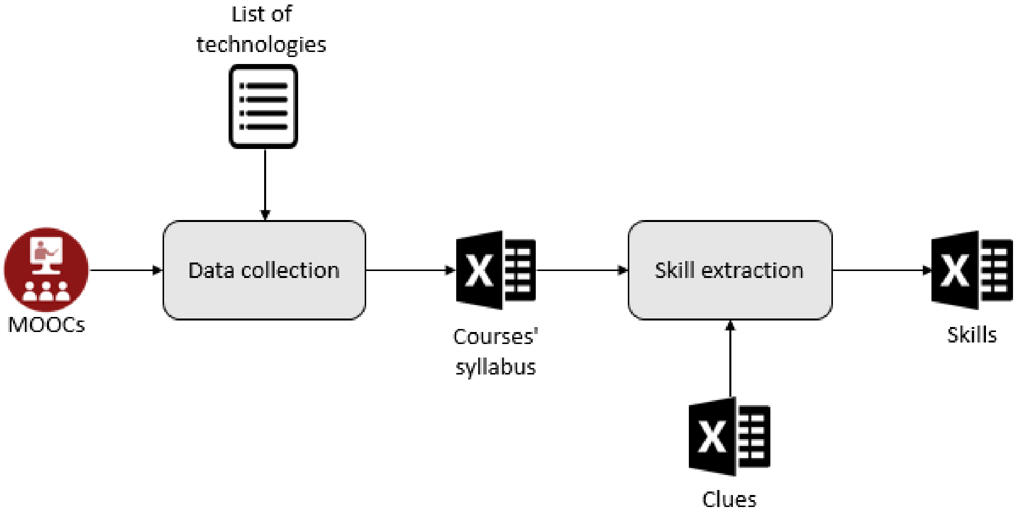


Figure 4 - Workflow for the extraction of skills from online courses (MOOCs)

5.1 Data collection

This first phase consists in automatically downloading, through web scraping[14], the content of online courses inherent to specific technologies (extracted from previous phase). The output of this first step is an excel file containing the description (syllabus) of every downloaded course which will be the input for the next step, i.e. skill extraction.

5.2 Skill extraction

There are two main inputs to this last phase: the *description* of courses inherent to specific technologies and a list of *clues*. The latter is a list in xlsx format containing words or strings (more words together) which, when found in a sentence, indicate with a high probability the presence of a skill. Example of clues are “Knowledge of”, “You will learn”, “plan”, “develop”, “use” etc. These clues are then fed to a program that goes through every course description searching and automatically extracting phrases containing these clues. The extracted phrases are then cleansed by keeping only that part of phrase that starts from the clue (e.g. “the user will learn machine learning”→ cleansing →”learn machine learning”).

An example of the results obtainable with this process are shown in *table 1,* where skills have been extracted from courses related to Cloud[15] technology found in Udemy[16].

[14] Web scraping, web harvesting, or web data extraction consists in the automatic extraction of content from websites.

[15] Cloud computing is the on-demand availability of computer system resources, like data storage and computing power, without direct active management by the user. Source: https://en.wikipedia.org/wiki/Cloud_computing

technology	course	Extracted skills
cloud computing	A quick introduction to Cloud computing	use Cloud platforms
cloud computing	A quick introduction to Cloud computing	Understand Virtualization and its use in Infrastructure development
cloud computing	A quick introduction to Cloud computing	build cloud infrastructure
cloud computing	Introduction to cloud computing	set up virtual servers
cloud computing	Introduction to cloud computing	work with cloud file storage
cloud computing	Introduction to cloud computing	use cloud collaboration options

Table 1 - Technologies extracted from courses related to Cloud technologies

6 Bridging to Job profiles

Finally, the last step consists in understanding which job profiles have knowledge on the extracted technologies. Not only is there a need to understand how technologies are changing and which skills will have a greater request in the future, but more importantly which occupations will be majorly impacted by such changes. To undertake this bridging phase, the developed approach exclusively refers to occupational frameworks. The latter are a reference point in labour market since they define key features of an occupation as a **standardized** and **measurable** set of variables.

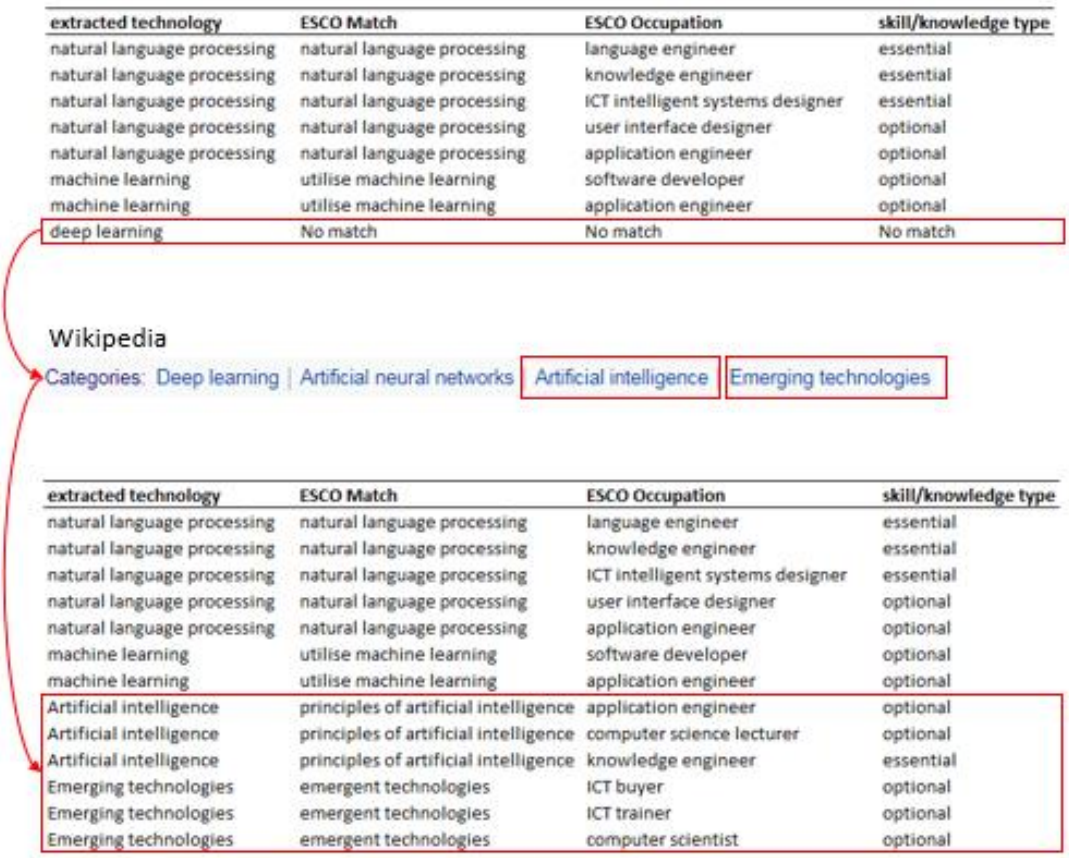
From a methodological point of view, this phase is carried out using the API (application programming interface) provided by the websites of these occupational frameworks, which gives the possibility to interact with the latter in a completely automatic way. Using ESCO's API[17] for instance, technologies previously identified are automatically matched to ESCO's correspondent skills or knowledge (ESCO does not refer to technologies) with a list of related job profiles and an indication of whether the skill/knowledge is essential or optional for that job profile. The program identifies a good matching through the calculation of a **similarity ratio**, which analyzes the literal similarity between the technology we want to match and the skill/knowledge that ESCO has returned as a match. The similarity ratio ranges between 0 and 1, and if its value is greater than or equal to 0.8 then we have a good matching with ESCO. Unfortunately, it can happen that we have a good match even though the ratio is lower than 0.8 (e.g. "artificial intelligence" and "principles of artificial intelligence" have a ratio of 0.76), for this reason, other than calculating the ratio, I also determine the *goodness*

[16] https://www.udemy.com/

[17] https://ec.europa.eu/esco/portal/api

of a match, if the extracted technology is **contained** within the ESCO match (e.g. "**artificial intelligence**" is contained in "principles of **artificial intelligence**").

Unfortunately, not all technologies find a direct match in ESCO. This can be due to the extracted technologies being more detailed than the ones present in the occupational frameworks. When this happens, I identify the above category of the extracted technology by searching for the related Wikipedia page and extracting what is found in the **category section** (the bottom part of every Wiki page). Matching is then repeated with the categories instead of the initial extracted technologies. A clear example of this process is shown in *figure 5.*



extracted technology	ESCO Match	ESCO Occupation	skill/knowledge type
natural language processing	natural language processing	language engineer	essential
natural language processing	natural language processing	knowledge engineer	essential
natural language processing	natural language processing	ICT intelligent systems designer	essential
natural language processing	natural language processing	user interface designer	optional
natural language processing	natural language processing	application engineer	optional
machine learning	utilise machine learning	software developer	optional
machine learning	utilise machine learning	application engineer	optional
deep learning	No match	No match	No match

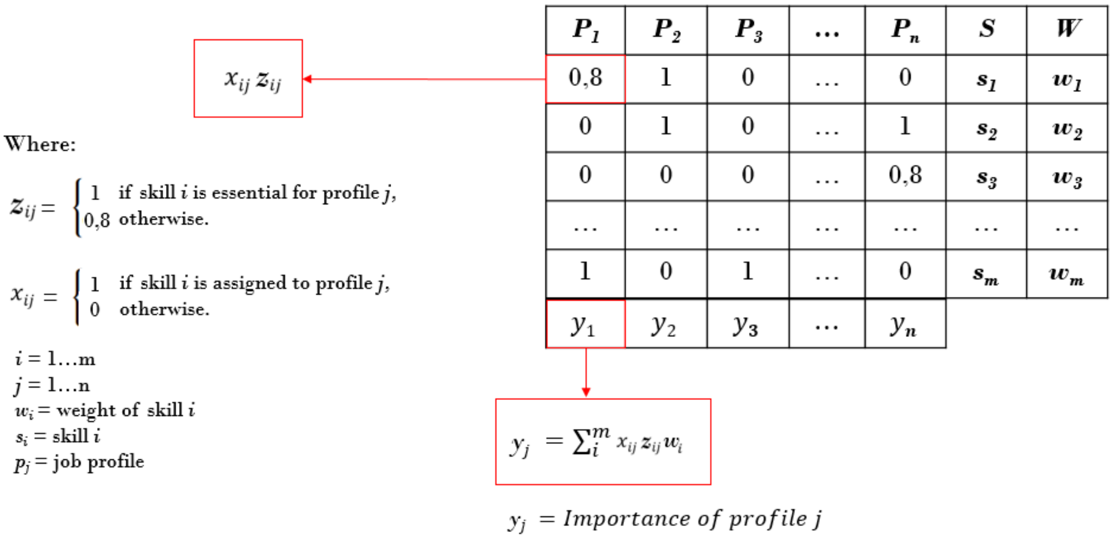
Wikipedia

Categories: Deep learning | Artificial neural networks | Artificial intelligence | Emerging technologies

extracted technology	ESCO Match	ESCO Occupation	skill/knowledge type
natural language processing	natural language processing	language engineer	essential
natural language processing	natural language processing	knowledge engineer	essential
natural language processing	natural language processing	ICT intelligent systems designer	essential
natural language processing	natural language processing	user interface designer	optional
natural language processing	natural language processing	application engineer	optional
machine learning	utilise machine learning	software developer	optional
machine learning	utilise machine learning	application engineer	optional
Artificial intelligence	principles of artificial intelligence	application engineer	optional
Artificial intelligence	principles of artificial intelligence	computer science lecturer	optional
Artificial intelligence	principles of artificial intelligence	knowledge engineer	essential
Emerging technologies	emergent technologies	ICT buyer	optional
Emerging technologies	emergent technologies	ICT trainer	optional
Emerging technologies	emergent technologies	computer scientist	optional

Figure 5 - ESCO matching

However, not all matched job profiles have the same relevance, for instance looking at results shown in *figure 5,* technologies are linked to more than one occupation. That being said, the question arises: Is it possible to give an order of importance to the matched occupations? Yes, the value of importance given to these profiles, strongly depends on the technology to which they are associated, because as previously stated, growth in technology translates into growth in skills and consequently job profiles. That said, the value is calculated by summing the occurrence of the technology the job profile is linked to multiplied by a coefficient whose value is 0.8 (determined through a sensitivity analysis), if the matched ESCO skill/knowledge is optional for the occupation, or 1 if it is essential. This process is outlined in *figure 6.*



$x_{ij}\, z_{ij}$

P_1	P_2	P_3	…	P_n	S	W
0,8	1	0	…	0	s_1	w_1
0	1	0	…	1	s_2	w_2
0	0	0	…	0,8	s_3	w_3
…	…	…	…	…	…	…
1	0	1	…	0	s_m	w_m
y_1	y_2	y_3	…	y_n		

Where:

$$z_{ij} = \begin{cases} 1 & \text{if skill } i \text{ is essential for profile } j, \\ 0{,}8 & \text{otherwise.} \end{cases}$$

$$x_{ij} = \begin{cases} 1 & \text{if skill } i \text{ is assigned to profile } j, \\ 0 & \text{otherwise.} \end{cases}$$

$i = 1 \ldots m$
$j = 1 \ldots n$
$w_i =$ weight of skill i
$s_i =$ skill i
$p_j =$ job profile

$$y_j = \sum_i^m x_{ij}\, z_{ij}\, w_i$$

$$y_j = Importance\ of\ profile\ j$$

Figure 6 - Determining the importance "y" of extracted job profiles

Once the relevance of the extracted job profiles has been calculated (automatically), it can then be visualized using a basic bar chart, where the length of each bar is proportional to the importance of the occupation. An example is shown below in *figure 7* where I have hypothesized, that the occurrence of Natural language processing, machine learning and deep learning in papers are respectively 25, 15 and 32*.*

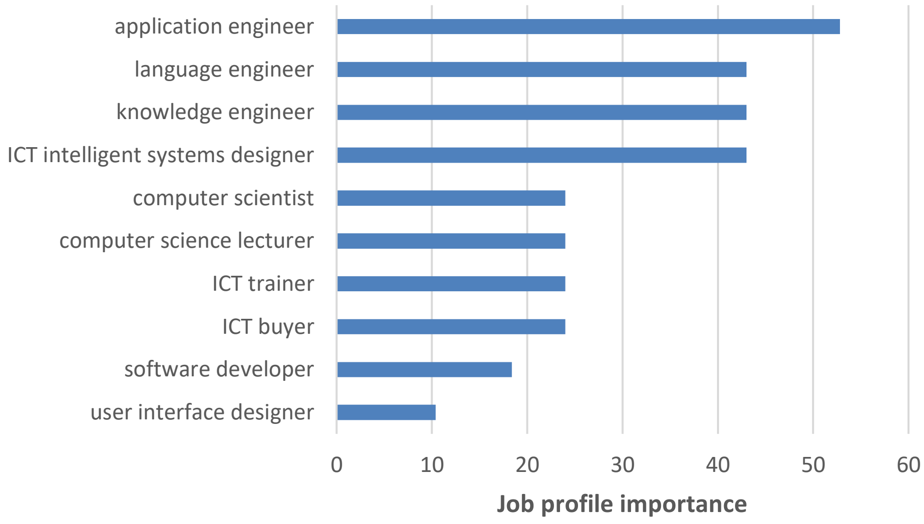


Figure 7 - Relevance of job profiles

From the results shown in *figure 7,* the *application engineer* is the most relevant occupation based on the occurrence of the technologies to which the profile is connected and thus will probably have a greater request tomorrow.

7 Final considerations

As seen, automated techniques can be very effective, and their use can provide valuable information on ongoing changes and future skill demand.

Unfortunately, the proposed approach in the present thesis, has unsolved issues. In detail:

- New job profiles cannot be detected as the link from technologies/skills to professional occupations is made through occupational frameworks, thus a pre-existing list;
- Future scenarios do not only depend on technological evolution, but there are also other drivers of change which might have an impact on occupations and skills (e.g. socioemotional skills) and have not been considered in this approach due to difficulty in extracting related information and processing them using automated techniques;
- ESCO, as seen, does not always find a match with extracted skills and technologies, thus not providing exact results. Perhaps, this issue could be solved by linking technologies or skills to job profiles found in global employment websites, i.e. extracting the name of the occupation addressed in job postings that contain the technology or skill of interest. However, even though this last process would work, it would not result in high value for stakeholders (e.g. employment services, training providers etc.) since job profiles addressed in vacancies are not expressed in a standard way as in occupational frameworks;
- The extraction of skills has to rely on the presence of courses, however, courses on a specific technology may not be available due to different factors, for instance the technology is too recent, thus courses have not been developed yet, or the use of certain technologies cannot be taught online;

In short, due to the above-mentioned limits, no methodology alone can rely solely on automated techniques. In fact, information which cannot be found in available or accessible data sources, could be obtained through experts' consultation, bearing in mind that objectivity may be undermined.

By ways of conclusion, the methodology that I have analysed must not be considered as a stand-alone tool, rather as a complementary set of tools to the more conventional approaches (surveys, focus groups, econometric models etc.).