Skills2Graph: Processing Million Job Ads to face the Job Skill Mismatch Problem

Anna Giabelli[1] , **Lorenzo Malandri**[2] , **Fabio Mercorio**[2] , **Mario Mezzanzanica**[2] , **Andrea Seveso**[1]

[1]Dept. of Informatics, Systems and Communication, University of Milano Bicocca, Milan, Italy
[2]Dept. of Statistics and Quantitative Methods, University of Milano Bicocca, Milan, Italy
anna.giabelli@unimib.it lorenzo.malandri@unimib.it fabio.mercorio@unimib.it
mario.mezzanzanica@unimib.it andrea.seveso@unimib.it

Abstract

In this paper, we present `skills2graph`, a tool that, starting from a set of users' professional skills, identifies the most suitable jobs as they emerge from a large corpus of 2.5M+ Online Job Vacancies (OJVs) posted in three different countries (the United Kingdom, France, and Germany). To this aim, we rely both on co-occurrence statistics - computing a count-based measure of skill-relevance named Revealed Comparative Advantage (*rca*) - and distributional semantics - generating several embeddings on the OJVs corpus and performing an intrinsic evaluation of their quality. Results, evaluated through a user study of 10 labor market experts, show a high P@3 for the recommendations provided by `skills2graph`, and a high nDCG (0.985 and 0.984 in a [0,1] range), that indicates a strong correlation between the experts' scores and the rankings generated by `skills2graph`.

1 Introduction and Motivation

Given the very high number of job positions and applicants on online job portals, the problem of person-job fit has become relevant in recent literature, both as a skill measuring system [Xu *et al.*, 2018] and as a job recommendation system [Zhang *et al.*, 2016]. Recommender systems in the labor market domain rely strongly on handcrafted features and expert knowledge, and these characteristics make them costly, difficult to update, and error-prone. For that reason, we propose `skills2graph`, a job recommendation system based on a knowledge-poor and data-driven approach, which can be adapted to different countries/industries and easily updated over time. `skills2graph` was realized as part of the research activity of an EU project[1], which aims at realizing the first EU real-time labor market monitor, by collecting and classifying Online Job Vacancies (OJVs) from all 27+1 EU countries, extracting the requested skills from job descriptions, see, e.g. [Boselli *et al.*, 2017; Boselli *et al.*, 2018a; Boselli *et al.*, 2018b; Colombo *et al.*, 2019; Giabelli *et al.*, 2020b]. `skills2graph` is designed as a tool to support decision making activities of citizens and labor market analysts facing the skill match/mismatch problem by processing 2.5M+ OJVs to answer the following questions:

Q1. **Skills-Occupation match.** How relevant are my skills for my current occupation?

Q2. **Reskilling.** Which are the relevant skills for my actual position that I should acquire?

Q3. **Career-path.** Which are the occupations that fit the skills I already have in a specific country? Which skills should I learn to have a better fit with one of those occupations, given its characteristics in that specific country?

Q4. **Upskilling.** Given a new skill that I want to acquire, what are other similar/complementary skills that I can get too?

The approach is based on the *rca* [Alabdulkareem *et al.*, 2018], a count-based measure that assesses the importance of a skill for an occupation based on their respective appearances in OJVs, and on the cosine similarity between the vectors representing skills and occupations in an embedding space. We generate several embeddings, and we perform an intrinsic evaluation of them, selecting the one which better expresses the semantic similarity between words within the European Standard Taxonomy of Occupations and Skills (ESCO[2]). This way, we produce a unified representation that preserves distributional and lexical-based semantic features.

Contribution. The contributions of `skills2graph` are:

(i) We exploit web labor market data using distributional semantics (embeddings), knowledge-based representations (ESCO), and a count-based measure of skill relevance (*rca*). We melt those representations to measure the person-job fit starting from the user's skills;

(ii) We organize the previous resources as a graph database for performing graph-traversal queries;

(iii) We present `skills2graph`: a recommender system that exploits the resources developed in (i) and (ii) to suggest the most suitable occupations starting from the user's skills in a certain context.

[1] This work is partially supported within the research activity of an EU Project entitled EUROSTAT 2020: *Towards the European Web Intelligence Hub - European system for collection and analysis of online job advertisement data (WIH-OJA)* in which some of the authors are involved as coordinators and researchers. https://tinyurl.com/eurostat-lm-2020

[2] https://tinyurl.com/sv4squr

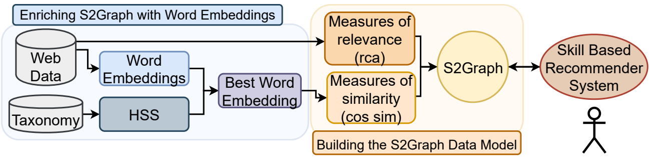

Enriching S2Graph with Word Embeddings
Web Data
Word Embeddings
Taxonomy
HSS
Best Word Embedding
Measures of relevance (rca)
Measures of similarity (cos sim)
S2Graph
Skill Based Recommender System
Building the S2Graph Data Model

Figure 1: Workflow of steps for building `skills2graph`

2 An Overview of `skills2graph`

The workflow of `skills2graph`, presented in Fig. 1, can be divided into four main steps.

S.1: Select a taxonomy-aware word vector representation. We train several embedding models through FastText and, to select the model which better preserves the semantic similarity between taxonomic elements, we perform an intrinsic evaluation following [Baroni *et al.*, 2014]. The evaluation consists of choosing the model which maximizes the correlation between cosine similarity between vectors and a benchmark measure of similarity. In [Baroni *et al.*, 2014] the authors use a human-crafted similarity dataset as the benchmark, however, those kinds of resources are costly to update, limited in size and coverage, and struggle to distinguish between semantic similarity and relatedness [Camacho-Collados *et al.*, 2017]. Hence, we employ an automated measure of pairwise semantic similarity between taxonomic elements, namely `HSS` (developed in [Giabelli *et al.*, 2021]).

Compared with `HSS`, previous measures of semantic similarity (see [Aouicha *et al.*, 2016]) suffer from two main limitations. First, when a word has multiple senses, those methods compute a value of similarity for each word sense and then consider only the highest, which is the self-information of the least frequent lowest common ancestor. As a result, more specific senses will have a higher value of similarity, but this does not reflect the use of words in OJVs. Second, though they consider the taxonomy's structure (i.e., the relationships between concepts), they do not acknowledge the number of child entities (i.e., words) belonging to those concepts. That is crucial in our case as ESCO includes generic concepts that contain many different occupations, while some specific concepts can be represented by a few occupations which are highly informative.

S.2: Compute the skill-occupation relevance. We compute the relevance of the skills for each occupation through the *rca*, originally used in the US context [Alabdulkareem *et al.*, 2018], where the authors used the measure *onet*, provided by the O*NET taxonomy[3], to take into account the importance of each skill for each occupation. ESCO does not provide such a measure, hence we employ the *skill frequency*, which for occupation $o_i \in \bar{O}$ and skill $s_l \in \bar{S}$ is defined as:

$$sf(o_i, s_l) = \frac{\sum_{k=1}^{m} I(o_k = o_i) \cdot I(s_i = s_l)}{\sum_{k=1}^{m} I(o_k = o_i)} \tag{1}$$

where I denotes the indicator function.

The *rca* is computed as:

$$rca(o_i, s_l) = \frac{sf(o_i, s_l) / \sum_{j=1}^{p} sf(o_i, s_j)}{\sum_{k=1}^{m} sf(o_k, s_l) / \sum_{k=1}^{m} \sum_{j=1}^{p} sf(o_k, s_j)} \tag{2}$$

where *sf* is the *skill frequency* of skill s_l for occupation o_i. In order to have a measure which is more easily understandable, we compute the *normalized rca*, normalizing the *rca* with respect to the maximum value obtained for the occupation taken into account:

$$rca_N(o_i, s_l) = \frac{rca(o_i, s_l)}{\max_j rca(o_i, s_j)} \tag{3}$$

so that $rca_N(o_i, s_l) \in [0, 1]$ and the most requested skill for each occupation has a *normalized rca* equals to 1.

S.3: Store data as a DAG. The information extracted through word embeddings and *rca* is stored in our graph database, called `S2JGraph`, which is formalized as a directed labeled multi-graph, and the formalization is inspired by [Giabelli *et al.*, 2020a; Mercorio *et al.*, 2019]. Notice both the *rca* and the best embedding are computed for each country, capturing the differences between the requested skills and the occupations as they are used in different countries.

S.4: Rank occupations. Given a starting set of skills $\mathbf{S}$, we rank the occupations on the basis of (i) the *cosine similarity* between the skills in $\mathbf{S}$ and the most required skills for occupation o_i and (ii) the *rca* of the skills in $\mathbf{S}$ for o_i:

$$rank_i = \frac{\sum_{s_j \in \mathbf{S}_i} rca_N(o_i, s_j) \cdot max_{s_l \in \mathbf{S}} [cos\_sim(s_j, s_l)]}{|\mathbf{S}_i|} \tag{4}$$

with $\mathbf{S}_i$ being the set of skills required by the occupation o_i with a *normalized rca* of at least 0.6.

3 Skill Based Recommendations

Having `S2JGraph` stored as a DAG in Neo4j allows to query it using the Cypher query language [Francis *et al.*, 2018]. `S2JGraph` is helpful for those who may want to know the importance of the skills they own for their actual occupation or other occupations, in their country or in a different one. They may also want to know which skill gap they have to fill

Skill $s \in \mathbf{S}$	$rca_N(c_S)$	$rca_N(c_A)$
implement front-end website design	0.59	0.60 ↔
CSS	0.51	0.63 ↑
C#	0.46	0.40 ↓
use markup languages	0.17	0.61 ↑↑

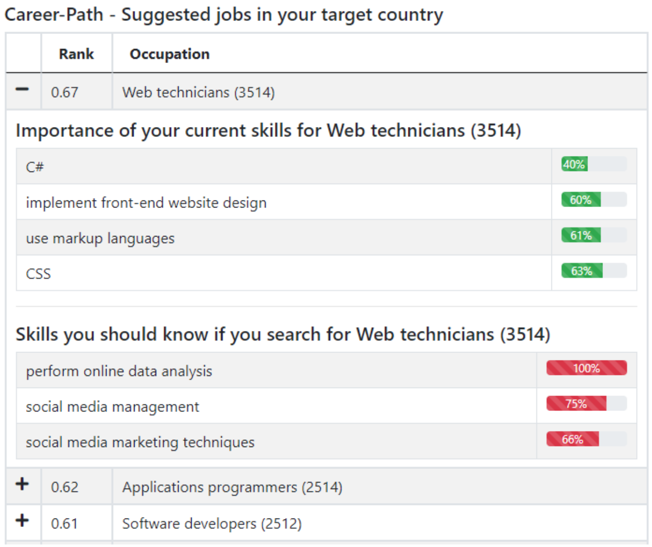
Skill gap for o_i	rca_N
perform online data analysis	1
social media management	0.75
social media marketing techniques	0.66

Table 1: Highest ranked result, "Web Technicians", showing the skills' rca_N in c_S and c_A and the skill gap o_i.

[3] https://www.onetonline.org/

	Starting skills	First Recommendation	Second Recommendation	Third Recommendation
1	IDE software, Agile project management, UML, ICT debugging tools	Software developers (0.37)	Software and applications developers and analysts NEC (0.23)	Applications programmers (0.16)
2	Web programming, Java, C++, Python	Software developers (0.28)	Web and multimedia developers (0.15)	Database and network professionals NEC (0.12)
3	Maintain database performance, Manage ICT data architecture, SQL, NoSQL	Database designers and administrators (0.41)	Systems administrators (0.19)	Software developers (0.16)

Table 2: Examples of three sets of starting skills and top recommendations for the user study.



Career-Path - Suggested jobs in your target country

	Rank	Occupation
−	0.67	Web technicians (3514)

Importance of your current skills for Web technicians (3514)

C#	40%
implement front-end website design	60%
use markup languages	61%
CSS	63%

Skills you should know if you search for Web technicians (3514)

perform online data analysis	100%
social media management	75%
social media marketing techniques	66%

+	0.62	Applications programmers (2514)
+	0.61	Software developers (2512)

Figure 2: A screenshot of skills2graph's interface, showing the Career-path suggestions for the example query.

to comply with the requirements of a specific occupation. Finally, they may want to know, given a new skill they would like to acquire if there are similar or complementary skills they might be interested in. We can say that S2JGraph helps the user in the problem of skill match/mismatch by answering the questions we draw in the Introduction, that are (**Q1**) Skills-Occupation match, (**Q2**) Reskilling, (**Q3**) Career-path, and (**Q4**) Upskilling.

The principal use case - **Q3** - recommends a series of occupations in a target labor market based on the user's skills, matching all the occupations in the target country c_A which require at least one of the starting skills in $\mathbf{S}$. Then the query matches all the skills which are required by the target occupation with a $rca > \alpha$ and which have a *cosine similarity* with all of the starting skills in $\mathbf{S} < \beta$. These are the skills of the *skill gap*, which are relevant for the target occupation (high rca) and different enough from the starting skills (low *cosine similarity*). These are skills that the user should acquire to do that job in the target country. An example of query **Q3** is reported in Tab. 1 ($\alpha = 0.6$, $\beta = 0.7$). The starting parameters are the following: $\mathbf{S}$ =[”implement front-end website design”, ”CSS”,”C#”, ”use markup languages”]; o_S =”Web and multimedia developers”; $c_S = UK$; $c_A = DE$. Fig. 2 shows the results in the user interface.

4 User Evaluation

The results of skills2graph were evaluated through a user study following [Kanakia *et al.*, 2019]. We asked ten labor Market experts belonging to the European Network on Regional Labour Market Monitoring to judge whether the starting skills are relevant for the occupations provided by **Q3**, using a Likert scale ranging from 1 to 5, with 1 being not relevant and 5 being completely relevant. The participants were all confident in their ability to correctly evaluating the recommendations, since they are active in the Labor Market Intelligence field, and well-acquainted in the ICT domain. The evaluation of skills2graph was performed on the British labor market, using ten different starting sets of four skills, three of which are shown in Tab. 2. The sets were chosen within the most popular skills in the ICT labor market, selecting the clusters with high similarity with each other - with a few changes when they were too similar. As three recommendations were presented for each item, we decided to use Precision@3 (P@3), assuming either a user score of at least 3 is a true positive (P@3-3) or of at least 4 (P@3-4). The normalized Discounted Cumulative Gain (nDCG) has also been computed, which measures the usefulness of an item based on its position in the list of recommendations.

5 Results and Discussion

All the ten experts responded to the user evaluation, and there were no missing votes. The results (see Tab. 3) show a high degree of correlation between the experts' evaluations and the recommendation ranking, showing that the system is effective in identifying the correct jobs given a set of user's skills. skills2graph can process any dataset of OJVs in any EU language; here we used 2.5M+ OJVs form France, Germany, and the United Kingdom, processed through distributional semantics and co-occurrence statistics, organized in a graph database. skills2graph identifies the most suited job based on a set of user's skills, encoding the skill relevance as it emerges from the labor market demand.

A **demo video** of the tool is available at https://youtu.be/Fiz9z_4FSbA.

	P@3-3	P@3-4	nDCG
Result	0.823	0.610	0.985

Table 3: User evaluation results for the two methods. P@3-N indicates a user score of at least N is considered a true positive.

References

[Alabdulkareem *et al.*, 2018] Ahmad Alabdulkareem, Morgan R. Frank, Lijun Sun, Bedoor AlShebli, César Hidalgo, and Iyad Rahwan. Unpacking the polarization of workplace skills. *Science Advances*, 4(7), 2018.

[Aouicha *et al.*, 2016] Mohamed Ben Aouicha, Mohamed Ali Hadj Taieb, and Abdelmajid Ben Hamadou. Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness. *Applied Intelligence*, 45(2):475–511, 2016.

[Baroni *et al.*, 2014] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*, 2014.

[Boselli *et al.*, 2017] Roberto Boselli, Mirko Cesarini, Fabio Mercorio, and Mario Mezzanzanica. Using machine learning for labour market intelligence. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18-22, 2017, Proceedings, Part III*, volume 10536 of *Lecture Notes in Computer Science*, pages 330–342. Springer, 2017.

[Boselli *et al.*, 2018a] Roberto Boselli, Mirko Cesarini, Stefania Marrara, Fabio Mercorio, Mario Mezzanzanica, Gabriella Pasi, and Marco Viviani. Wolmis: a labor market intelligence system for classifying web job vacancies. *J. Intell. Inf. Syst.*, 51(3):477–502, 2018.

[Boselli *et al.*, 2018b] Roberto Boselli, Mirko Cesarini, Fabio Mercorio, and Mario Mezzanzanica. Classifying online job advertisements through machine learning. *Future Gener. Comput. Syst.*, 86:319–328, 2018.

[Camacho-Collados *et al.*, 2017] Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26, 2017.

[Colombo *et al.*, 2019] Emilio Colombo, Fabio Mercorio, and Mario Mezzanzanica. Ai meets labor market: Exploring the link between automation and skills. *Information Economics and Policy*, 47, 2019.

[Francis *et al.*, 2018] Nadime Francis, Alastair Green, Paolo Guagliardo, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Stefan Plantikow, Mats Rydberg, Petra Selmer, and Andrés Taylor. Cypher: An evolving query language for property graphs. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1433–1445. ACM, 2018.

[Giabelli *et al.*, 2020a] Anna Giabelli, Lorenzo Malandri, Fabio Mercorio, and Mario Mezzanzanica. Graphlmi: A data driven system for exploring labor market information through graph databases. *Multimedia Tools and Applications*, pages 1–30, 2020.

[Giabelli *et al.*, 2020b] Anna Giabelli, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Andrea Seveso. NEO: A tool for taxonomy enrichment with new emerging occupations. In *International Semantic Web Conference*, pages 568–584. Springer, 2020.

[Giabelli *et al.*, 2021] Anna Giabelli, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Andrea Seveso. Skills2job: A recommender system that encodes job offer embeddings on graph databases. *Applied Soft Computing*, 101:107049, 2021.

[Kanakia *et al.*, 2019] Anshul Kanakia, Zhihong Shen, Darrin Eide, and Kuansan Wang. A scalable hybrid research paper recommender system for microsoft academic. In *WWW*, pages 2893–2899, 2019.

[Mercorio *et al.*, 2019] Fabio Mercorio, Mario Mezzanzanica, Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperlì. A tool for researchers: Querying big scholarly data through graph databases. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part III*, volume 11908 of *Lecture Notes in Computer Science*, pages 760–763. Springer, 2019.

[Xu *et al.*, 2018] Tong Xu, Hengshu Zhu, Chen Zhu, Pan Li, and Hui Xiong. Measuring the popularity of job skills in recruitment market: A multi-criteria approach. In *AAAI*, 2018.

[Zhang *et al.*, 2016] XianXing Zhang, Yitong Zhou, Yiming Ma, Bee-Chung Chen, Liang Zhang, and Deepak Agarwal. Glmix: Generalized linear mixed models for large-scale response prediction. In *SIGKDD*, pages 363–372, 2016.