SKILLS2JOB: A RECOMMENDER SYSTEM THAT ENCODES JOB OFFER EMBEDDINGS ON GRAPH DATABASES

A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, A. Seveso

University of Milano-Bicocca & CRISP Research Center - Italy

Introduction and Contribution

We propose a recommender system that, starting from a set of users' skills, identifies the most suitable jobs as they emerge from a large dataset of Online Job Vacancies (OJVs). To this aim, we process 2.5M+ OJVs posted in three different countries (United Kingdom, France, and Germany). The contribution of this paper is threefold:

(i) We exploit web labor market data using distributional semantics (embeddings), knowledge-based representations (ESCO), and a count-based measure of skill relevance (*rca*);

(ii) We organize the above-mentioned resources as a graph database;

(iii) We present `skills2job` , a recommendation 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.

The `skills2job` Approach

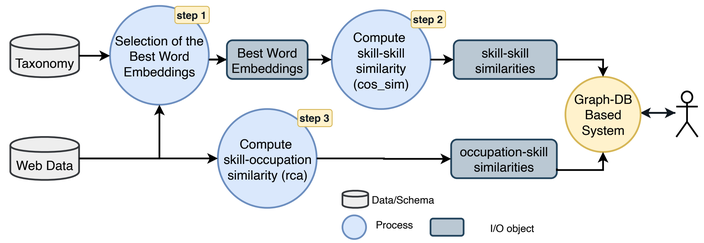

Fig. 1: A graphical overview of TaxoRef

Embeddings Evaluation

We generate several vector representations of the OJVs' text corpus through FastText. We **select the one that the maximizes the correlation between cosine similarity and taxonomic (ESCO) similarity**. The taxonomic similarity is expressed by a novel measure named Hierarchical Semantic Similarity (`HSS`):

$$sim_{HSS}(w_1, w_2) = \sum_{\ell \in \mathcal{L}} \hat{p}\left(\ell = L \mid w_1, w_2\right) \times IC(L) \quad (1)$$

Revealed Comparative Advantage

The **Revealed Comparative Advantage** (**rca**) was used in 2018 to assess the relevance of skills concerning occupations in the US context. The *rca* for o_i and s_l is defined as:

$$rca(o_i, s_l) = \frac{sf(o_i, s_l) / \sum_{j=1}^{p} sf(o_i, s_j)}{\sum_{k=1}^{m} sf(o_k, s_l) / \sum_{k=1}^{m} \sum_{j=1}^{p} sf(o_k, s_j)} \in [0, +\infty) \quad (2)$$

where *sf* is the **skill frequency** $sf(o_i, s_l) = \frac{\sum_{k=1}^{m} I(o_k=o_i) \cdot I(s_i=s_l)}{\sum_{k=1}^{m} I(o_k=o_i)}$. To have a measure more easily understandable, we created the **normalized rca**, normalizing with the maximum value of *rca* for the occupation.

Graph Database

`skills2job` uses the **graph database** `S2JGraph` as a convenient way of storing the labor market data to use them for recommending jobs.
The `S2JGraph` data model is represented in Fig. 2, where the **skill-similarity** is the **cosine similarity** between the vectors representing the skills in the best embedding model, while the **skill-requested** value is the **rca** of the skill for the occupation.

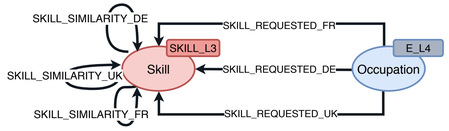

Fig. 2: Data Model of `S2JGraph`

Skill Based Recommendations

Given a set of starting skills $\mathbf{S}$, a starting occupation o_S, a starting country c_S, an arriving country c_A and a target skill s_T, all provided by the user, `skills2job` gives back to him/her:

(i) The relevance of each $s \in \mathbf{S}$ for o_S in c_S;

(ii) A list of occupations $\mathbf{O}$ in c_A and for each $o_i \in \mathbf{O}$:

- The indication of the relevance of each $s \in \mathbf{S}$ with respect to o_i;
- A list of skills that o_i requires and that are relevant for it and different from those in $\mathbf{S}$ (*gap skills*).

(iii) A list of skills recommended to the user because of $\mathbf{S}$ and s_T.

Experimental Results

The first part of the second and main query is shown in Tab. 1 in which are listed the first three recommendations in Germany. For each recommendation, the starting skills required by that occupation are shown, with their corresponding *normalized rca*.

Rank	Arriving occupation	4 skills	rca_{NORM}
0.56	Web Technicians	C#	0.3996
		implement front-end website design	0.5967
		use markup languages	0.6066
		CSS	0.6264
0.2	Applications programmers	C#	0.3293
		implement front-end website design	0.143
		use markup languages	0.1832
		CSS	0.1265
0.18	Software developers	C#	0.3145
		implement front-end website design	0.1327
		use markup languages	0.1614
		CSS	0.130

Tab. 1: Example of query (ii) with $c_S = UK$ and $c_A = DE$.

The `skills2job`'s recommendations were evaluated through a user study: we asked ten labor market experts to judge whether the starting skills are relevant for the occupations provided by the system or not. The evaluation of `skills2job` was performed on the British labor market, using ten different starting sets of four skills.
In Tab. 2 ere are shown the results for P@3-3, P@3-4, and nDCG for the two methods used to perform the task (ii).

	rcaB	**cosB**
P@3-3	0.823	0.763
P@3-4	0.610	0.570
nDCG	0.985	0.984

Tab. 2: User evaluation results for the two methods. P@3-N indicates that a user score of at least N is considered a true positive.

The first method (rcaB) lets us rank the occupations based on the rca_{NORM} with which the occupations require the starting skills. The second method (cosB) lets us rank the occupations based on the *cosine similarity* between the starting skills in $\mathbf{S}$ and the most required skills for o_i.
rcaB outperforms cosB in precision, despite both the methods obtained good results in P@3 and nDCG. The nDCG scores for both methods are similar and close to 1. These results suggest that there is a high degree of correlation between the user evaluation and the ordering rank of our recommendations.