Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 162 (2019) 857–864

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Information Technology and Quantitative Management (ITQM 2019)

An automatic skills standardization method based on subject expert knowledge extraction and semantic matching

Juan Bernabé-Moreno[a], Álvaro Tejeda-Lorente[a], Julio Herce-Zelaya[a], Carlos Porcel[b], Enrique Herrera-Viedma[a,1]

[a]*Department of Computer Science and A.I., University of Granada, Granada E-18071, Spain*
[b]*Department of Computer Science, University of Jaén, Jaén E-23071, Spain*

Abstract

The job market is rapidly changing. Artificial Intelligence and automation technologies are reshaping the career market. Everyday, new jobs appear and new skills are added to the scope of existing job profiles. At the same time, some skills that once were assumed to be "must-haves" for particular jobs are no longer requested and some jobs are even becoming obsolete. The speed of changes as well as the increasing complexity of the job market introduce a key new challenge: there is no clear definition for a particular job in terms of skills and scope and consequently, people holding the same job title cannot be assumed to be actually doing the same thing. In addition, applicants find difficult to develop career paths, as the mapping of skills to particular jobs are fuzzier than ever before. In this article, we present a novel approach to homogenize the job definition, gathering first subject matter expertise using semantic expansion techniques on collaborative wikies, applying a word embeddings supported method to mine the skills from existing job posts and finally executing a semantic matching algorithm to converge to a consistent skills mapping. In order to show how our method performs, we apply it to one of the most popular, yet heterogeneous modern jobs, the *data scientist* and discuss the results obtained for the English speaking market.

Keywords: collaborative wikies, word embeddings, semantic matching, job market, machine learning, skills modelling

1. Introduction

One of the most impacted areas by the increasing digitalization is the job market. New technologies, such as artificial intelligence, cognitive algorithms, etc., can be applied to perform certain tasks in a more reliable, more efficient and more cost-effective way than humans. The job market is being therefore continuously disrupted [1]. We witness new jobs emerging almost every day, existing jobs being consolidated into new more cross-discipline ones, jobs disappearing, jobs coming back after certain time being extinct. According to Bakhshi et al. [2], by 2030 it is likely that 20% of all occupations will be cut down while 10% will grow, having all the remaining occupations an uncertain outlook.

*Corresponding author.
E-mail address: viedma@decsai.ugr.es.

1877-0509
10.1016/j.procs.2019.12.060

In this frenetic environment, a new issue emerges: it is increasingly difficult to develop the same understanding for the scope of a particular job and therefore the skills required to perform it. If we take for example “baker" or “butcher", we can easily list the tasks and the skills required to perform the job. If we now consider the job “growth hacking expert", and ask several people, we are going to have substantially more diverging answers. Moreover, as “growth hacking expert" is a profile that usually exists in several industries, making it industry specific might contribute to broaden the understanding about scope or tasks and skills (e.g.: “growth hacking expert for pharma", “growth hacking expert for banking", etc.)

As a consequence, jobs labelled with the same title are less and less comparable on one hand, and educational/career path struggle to define the required skills to qualify for a particular job on the other hand, as these skills can be very diverse and changing over time. Not only individuals are affected, policy-makers, businesses and educational institutions need to react quicker than ever before to adapt and prepare the existing and future workforce for emerging requirements.

In this paper, we suggest a novel approach to automatically extract, augment and standardize the skills and scope for a particular job title. Our method relies on a well structured pipeline combining several NLP, clustering, deep-learning and semantic technologies, each technology to deliver optimized results for the different sub-tasks. To our knowledge, there is no other method proposed in the literature so far, to solve the skills standardization issue with the same degree of automation and quality, as our approach:

- extracts all relevant named entities corresponding to skills and tasks from all variety of jobs posts for a particular title (e.g.: "growth hacker")
- applies word embeddings to mine the relationships between the skills and tasks and extracts the most relevant groups applying density base clustering on the embeddings space
- mines the domain knowledge expertise for the extracted entities using collaborative wikies, enabling the semantic expansion of concepts and categories to build up a skills graph
- enables the convergence to a homogeneous set of entities applying structural and dynamic properties on the skills graph as well as semantic matching with previously identified entities

In this paper, after reviewing the supporting research background, we will present our method and explain the different components of the system to implement it. After that, we discuss the results of applying our method to standardize one of the most difficult and broadest jobs of the last decade: "data scientist", showing the impact of different parametrizations over a set of 17k jobs posts from different jobs portals. We finalize the paper providing the concluding remarks and pointing to further research lines.

2. Background

Skills modelling is a problem that has been addressed in different research fields for a long time. When the semantic web technoligies emerged, different ontologies were proposed to model competences and skills. In [3], the authors describe the use of ontologies as the enabling semantic infrastructure of competency management in a corporation, focusing on how it is supporting the knowledge creation cycle in the context of competency definitions. In [4] a methology for application-driven development of skills ontologies is presented as the core of a knowledge management system in a Swiss company.

Ontology based approaches are easy to understand but present three major disadvantages in rapidly changing environment: the need for a central authority to consolidate and curate the ontology, the rigidity to accommodate rapid changes and the amount of work required to generate different versions for different verticals, segments, industries. In addition, it is difficult to track history of changes over time. In spite of these issues, 2 major skills, competences and occupations networks are still maintained by experts and extensively used (mainly for research purposes): O*NET (Occupational Information Network [5]) [1] and ESCO (European Skills, Competences, Qualifications and Occupational Information Network [6]) [2]

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

[2] https://ec.europa.eu/esco/portal/home

More automatic approaches were pursued for skills extraction and modeling. In [7], Kivimaki et al. suggest a system to extract the relevant skills given an input text using a document similarity based approach combining manually extracted skills from LinkedIn and the Wikipedia hyperlink graph. The authors of SKILL [8] proposed a system for skills discovery and normalization in 2 modules: one to automatically generate a skill taxonomy (built upon skill related sections of resumes and Wikipedia categories) and one for skills tagging, which leverages properties of semantic word vectors to recognize and normalize relevant skills in input text.

Djumalieva and her coauthors [9] proposed a taxonomy extraction method based on jobs adverts applying network community detection algorithms and measuring the strength of the match skills-job with a co-occurrence frequency weighting logic.

3. An automatic skills standardization method

Our method consists of a well defined pipeline of different modules (see Fig. 1). Each module is designed to solve a particular task and leverages the most advanced technology for this purpose. In this section, we will describe the different modules:

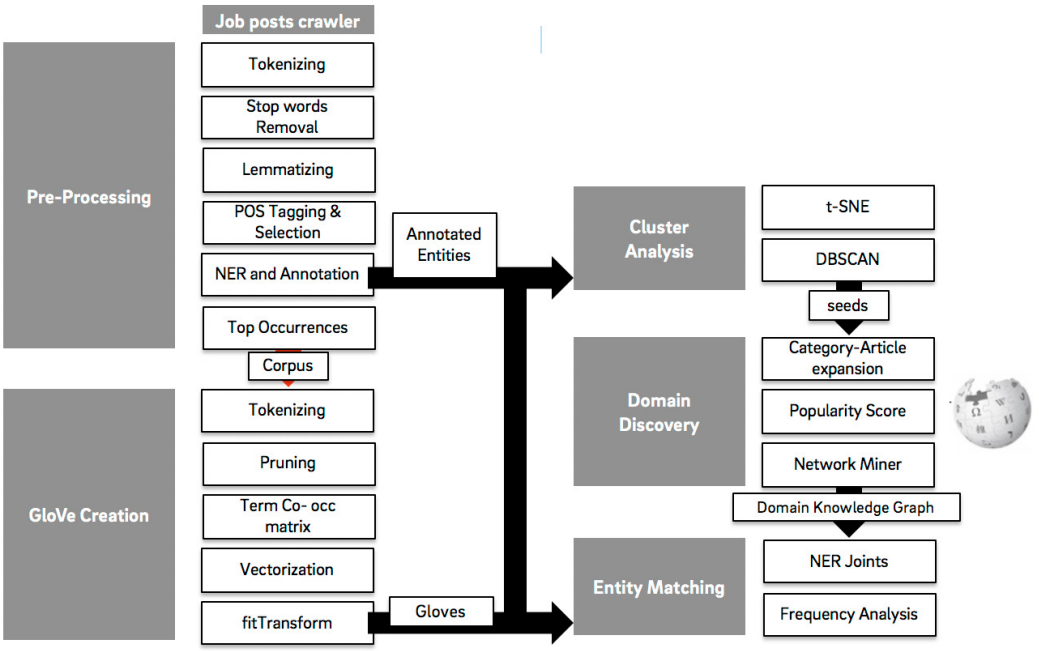

Fig. 1. System Overview

3.0.1. Jobs Posts Crawler

The **jobs posts crawler** connects to different job portals to perform queries to retrieve all jobs posts for a particular title (e.g.: “growth hacker"). In addition, further filtering possibilities are implemented to enable for standardization of the job in different context (e.g.: industry, geographical region, etc.) as well as adjusting the period of time within the jobs are crawled (it also allows for comparison of different periods to for example understand how the skills have changed over time for the same job). The result is a collection of job titles and their corresponding free-text description of scope (tasks and responsibilities) and of skills or requirements.

3.0.2. Pre-processing

The purpose of the pre-processing phase is two folded. Taking as input the collection of harvested jobs description from the previous step, it annotates the most relevant entities on one hand and generates a normalized

corpus where these entities are present for the embeddings extraction on the other hand. The pre-processing takes places applying tokenization [10], removal of stop words [11], lemmatization and Part of Speech tagging (implemented with [12]) and selection of particular PoS tags (nouns, verbs, adjectives) and filtering by a minimum of occurrences (to avoid sparsity and noise).

3.0.3. GloVe Creation

GloVes are global vectors capturing the semantic relationships between terms. To obtain the GloVes, we apply the algorithm described in [13] to the standardized corpus we created in the previous step. The GloVes provide the proper representation to perform algebraic operations on terms, for example computing the cosine distance between terms, required for further steps in the whole process.

3.0.4. Cluster Analysis

Once we have a) all annotated entities and b) their distances we can apply t-SNE [14] and DBSCAN [15] or any other density based clustering algorithm to identify the set of clusters grouping the entities with the highest similarity (the shortest distance). The centroids of the cluster can be used as seed entities, as required in the subsequent step.

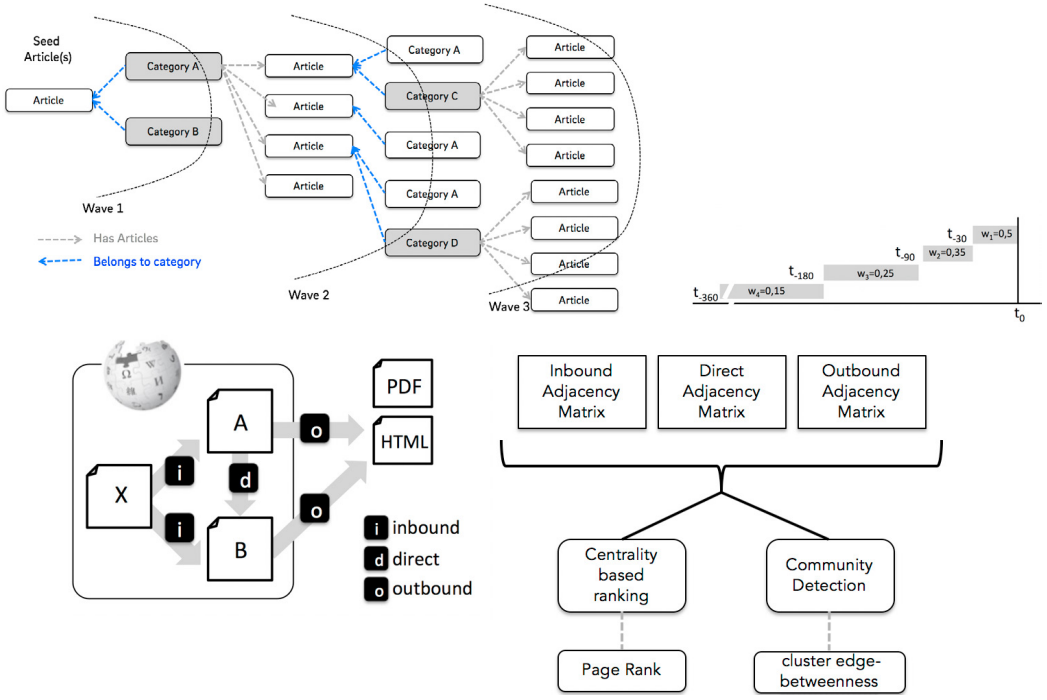

Fig. 2. a) Discovery Waves concept, b) Differential weighting for discovery periods and c) weighting according to network structure as explained in [16]

3.0.5. Domain Discovery

After obtaining the seed entities as defined in [16] in the previous step, we perform the domain discovery in waves (see Fig. 2 a), we rank the entities based on their popularity score (using a time window size weighting schema, as shown in Fig. 2 b) and we finally adjust the ranking applying the weights computed using the logic specified in the network miner of the same publication (see Fig. 2 c)) The result of this phase is the semantic network representing all the domain knowledge related to skills provided as seed in the previous steps after applying semantic expansion on the Wikipedia articles graph and skills ranking following the method we presented in our previous work.

3.0.6. Entity Matching

The final step is to consolidate the results by applying semantic matching using NER joints with the entities obtained in the previous phase. It is possible to relax the matching by allowing the next level of linked skills in the skills graph obtained in the previous step.

4. Experimental results and discussion

To show how our method works, we focused on one of the most popular, yet most difficult to specified jobs present: *data scientist*, due to following reasons: a) it's very actual and recent, b) it is considered one of the sexiest jobs in the 21st century, c) it is cross-industry, d) it is very broad in scope and unclear in the definition.

We set up the harvester using the API of well known international jobs portal to download 17.5K "data scientist" jobs posts starting between Oct 2018 and Dec 2018, all located in the largest English speaking countries (USA, Canada, Australia and UK). The Fig. 3 shows the considerable variety present already in the job titles. Using the *Google Cloud Entity POS Tagger* [3], we annotated all entities present in the text and extracted only the categories defining skills and tasks (nouns and verbs).

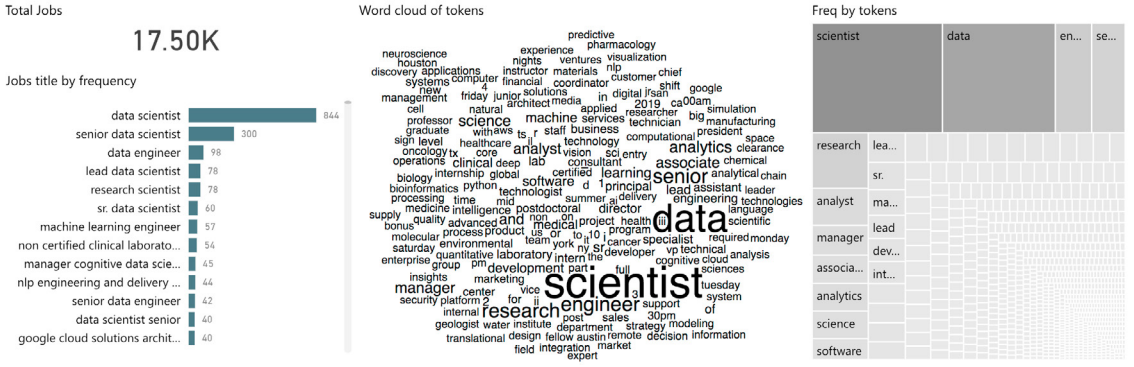

Fig. 3. Data Scientists jobs overview: a) Variety of jobs titles, b) Word-cloud with main terms and c) Top most frequent terms

After having the standardized and curated corpus, we applied the GloVe algorithm ([13]) to obtain the global vectors with a word vector size of 50 and a *fitTransform* phase running over 10 iterations (the choice of these values might minimally change the resulting vectors, but it shouldn't impact the overall result of our method). The embeddings projector contains more than 12K terms and can be visualized under following link: `https://bit.ly/2Zwxfyt`. To provide a tangible example, we show in Fig. 4 the embeddings cloud with the closest terms for "Python" as well as the distances to the nearest terms.

The cosine distance is then computed and density clustering is applied. The two leading DBSCAN parameters (*minPts* and *eps*) need tuning to yield a high number of clusters we can use in the next phase. Alternatively, it is possible to use a more advanced density based clustering algorithm, such as OPTICS ([17]), where just the *minPts* has to be adjusted. We worked with biggest 58 clusters. The tool provided in `https://bit.ly/2Zwxfyt` allows for testing different clustering methods (t-SNE, PCA, etc).

The domain discovery phase resulting in a massive semantic and skills network expansion. Due to the extension of the skills graph (over 20K entries), we will further show the application of our method focusing on a particular skills area, common to all data scientists jobs: "programming languages" (see Fig. 5). The semantic expansion procedure identified 5133 different articles grouped in 233 categories. Fig. 5 shows the closest terms to "Python", that happen to be the name of further programming languages, such as "Java", "R", "Scala", "C", "SQL" or "Matlab" (which intuitively confirms the power of the embeddings-based modelling).

The entity matching for "Python" is shown in Fig. 6: the green nodes are the ones remaining after the matching step (vs. all other nodes in black resulting from the domain discovery phase). We have created a video to show

[3] `https://cloud.google.com/natural-language/docs/analyzing-entities`

for all the Python-related skills -typical requirement for modern data scientists- that we have extracted applying semantic expansion (as explained in the subsection 3.0.5), how the semantic matching restricts the choice to the ones really in demand in the job markets `https://youtu.be/h08LqjUAT1c`

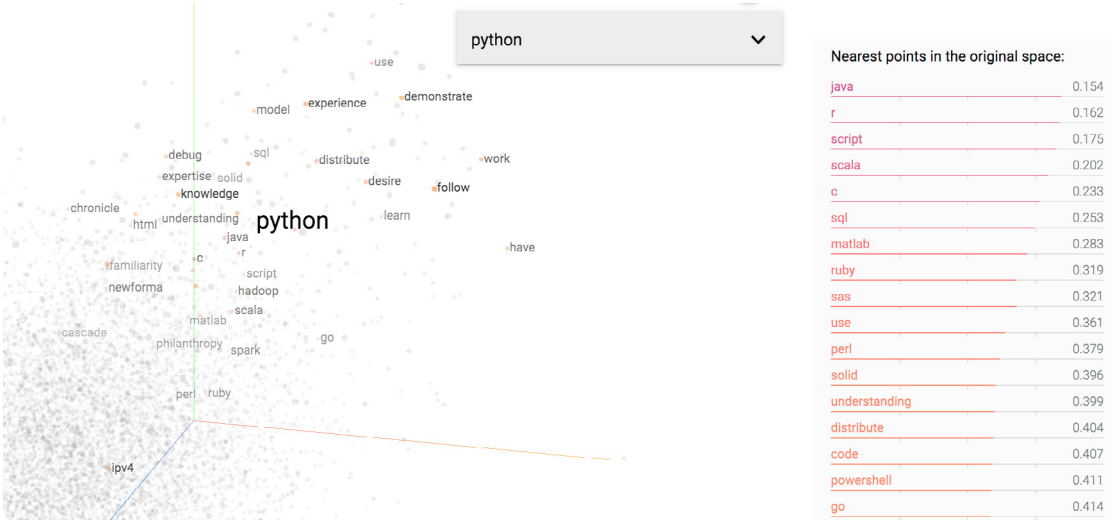

Fig. 4. a) Visualization of the closest terms for “Python" in the embeddings cloud and b) the distances to other terms

There is no ground truth or test set to validate our approach. Moreover, there is no similar method to our knowledge to standardize the homogenization of skills, yet the quality and validity of our proposal can be derived from the approach itself:

- The methodology we are proposing in this article is well defined and the result of each step can be easily reproduced.
- The quality of the NER annotation depends on the implementing technology but the current state of the art is delivering astonishing results (see the benchmark presented in [18] comparing the latest deep learning based NER methods).
- The GloVe algorithm can be made deterministic by setting the same pre-defined seed, preventing a parallel execution and restricting the number of workers to 1.
- The newest developments of density based clustering algorithms [19] and [20] while relying in highly effective clustering techniques provide full coverage over the occurrences space.
- The approach to holistically model the expert knowledge presented in[16] ensures a proper categorization of the most relevant skills by popularity, trending nature and positioning in the skills graph. In addition and as shown in the provided video (`https://youtu.be/h08LqjUAT1c`), it is possible to enlarge the number of skills accepted as standard for a particular job by modifying the parameters used to build the skills graph (centricity, etc). It gives the user the choice about how broad or specific a job post shall be.
- The semantic matching ensures that our method converges to the most relevant skills for any job.

5. Concluding remarks

In this paper we proposed a novel method to standardize in an automatic way the extraction of the skills and tasks required for a particular job. The input of our method is a set of job posts corresponding to a particular search for a job title. After some processing, the named entities are annotated and extracted into a corpus. From this corpus, embeddings for all terms in the corpus are extracted applying the Global Vectors algorithm. The distance between each vector is computed and on top a density cluster algorithm is applied to determine which entities are the ones with the highest weight (centroids). These entities are the most representative for the original

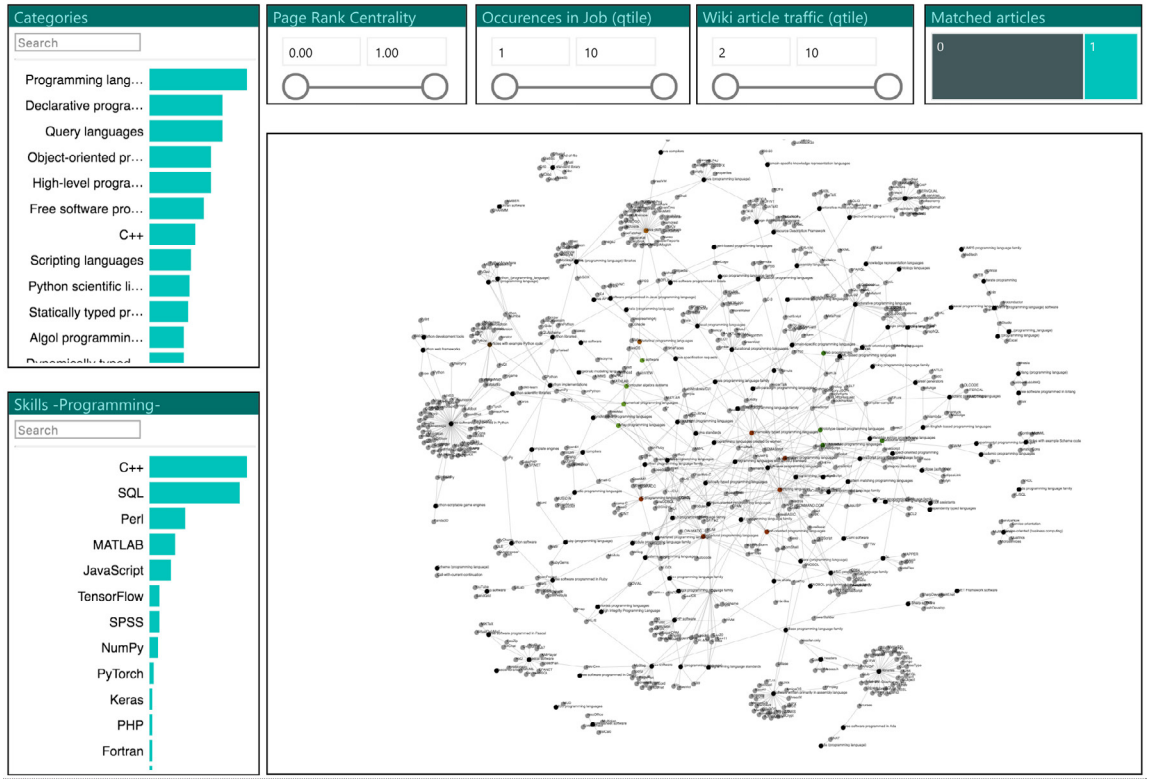

Fig. 5. Skills inspection dashboard focused on “programming languages"

job description and will serve as seeds for the collaborative wikies based semantic expansion algorithm described in [16]. The result is a ranked knowledge graph representing all significant skills and tasks a subject matter experts doing the aforementioned job is supposed to master. Last step is a pruning of the knowledge graph performed by semantic entity matching with the original set of annotated entities. We have implemented the system using 17.5K “data scientist" jobs posts and discussed the results and the performance.

Further research will focus on the analysis of changes in the scope and skills required for a particular job over time. Quantifying the degree of expertise and the experience required to develop a particular skill (junior, senior, etc.) can be explored using fuzzy linguistic modelling, which can also be used to determine to which degree the skills of a particular job can be automated. This could then be the basis to compute to which degree a particular job can be done by an artificial intelligence system.

6. Acknowledgments

This paper has been developed with the FEDER financing under Project TIN2016-75850-R

References

[1] J. E. Stiglitz, Ai, worker-replacing technological change and income distribution.
[2] H. Bakhshi, J. M. Downing, M. A. Osborne, P. Schneider, The future of skills: Employment in 2030, Pearson London, 2017.
[3] M.-A. Sicilia, Ontology-based competency management: infrastructures for the knowledge intensive learning organization, in: Intelligent learning infrastructure for knowledge intensive organizations: A semantic Web Perspective, IGI Global, 2005, pp. 302–324.
[4] T. Lau, Y. Sure, Introducing ontology-based skills management at a large insurance company, in: Proceedings of the Modellierung, Citeseer, 2002, pp. 123–134.
[5] N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret, E. A. Fleishman, Development of prototype occupational information network (o* net) content model. volume i: Report - volume ii: Appendices.

Fig. 6. Skills graph for “Python" after semantic expansion and remaining entities after semantic matching (in green)

[6] J. Smedt, M. Vrang, A. Papantoniou, Esco: Towards a semantic web for the european labor market 1409.

[7] I. Kivimäki, A. Panchenko, A. Dessy, D. Verdegem, P. Francq, H. Bersini, M. Saerens, A graph-based approach to skill extraction from text, in: Proceedings of TextGraphs-8 Graph-based Methods for Natural Language Processing, 2013, pp. 79–87.

[8] M. Zhao, F. Javed, F. Jacob, M. McNair, Skill: A system for skill identification and normalization, in: 27th IAAI Conference, 2015.

[9] J. Djumalieva, C. Sleeman, An open and data-driven taxonomy of skills extracted from online job adverts, Developing Skills in a Changing World of Work: Concepts, Measurement and Data Applied in Regional and Local Labour Market Monitoring Across Europe (2018) 425.

[10] R. Dridan, S. Oepen, Tokenization: Returning to a long solved problem—a survey, contrastive experiment, recommendations, and toolkit—, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vol. 2, 2012, pp. 378–382.

[11] A. Schofield, M. Magnusson, D. Mimno, Pulling out the stops: Rethinking stopword removal for topic models, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2017, pp. 432–436.

[12] H. Schmid, Improvements in part-of-speech tagging with an application to german, in: In proceedings of the ACL Sigdat-Workshop, Citeseer, 1995.

[13] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[14] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (Nov) (2008) 2579–2605.

[15] B. Borah, D. Bhattacharyya, An improved sampling-based dbscan for large spatial databases, in: International Conference on Intelligent Sensing and Information Processing, IEEE, 2004, pp. 92–96.

[16] J. Bernabé-Moreno, Á. Tejeda-Lorente, C. Porcel, E. Herrera-Viedma, A holistic domain knowledge discovery and recommendation system for collaborative wikis, in: SoMeT, 2017.

[17] M. Ankerst, M. M. Breunig, H.-P. Kriegel, J. Sander, Optics: ordering points to identify the clustering structure, in: ACM Sigmod record, Vol. 28, ACM, 1999, pp. 49–60.

[18] V. Yadav, S. Bethard, A survey on recent advances in named entity recognition from deep learning models, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 2145–2158.

[19] M. Alswaitti, M. Albughdadi, N. A. M. Isa, Density-based particle swarm optimization algorithm for data clustering, Expert Systems with Applications 91 (2018) 170–186.

[20] L. C. C. Heredia, A. R. Mor, Density-based clustering methods for unsupervised separation of partial discharge sources, International Journal of Electrical Power & Energy Systems 107 (2019) 224–230.