Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 00 (2017) 000–000

Procedia
Computer Science

www.elsevier.com/locate/procedia

Information Technology and Quantitative Management (ITQM 2017)

Analytics in Human Resource Management The OpenSKIMR Approach

Peter Mirski[a]*, Reinhard Bernsteiner[a], Dania Radi[a]

[a]Management Center Innsbruck (MCI), Department Management, Communication & IT, Universitaetsstrasse 15, 6020 Innsbruck, Austria

Abstract

Matching skill sets of individuals with highly demanded skill sets of jobs or occupations in the IT area is a great challenge – adding necessary learning items and visualizing the result is a very promising end to end approach. With the "Open Skill Match Maker" (OpenSKIMR) young people shall be able to plan and simulate their individual learning and career routes towards their desired career destination, like with classical route planning software. Using the ESCO, a multilingual classification of occupations, skills, competences and qualifications, will ensure a consistent understanding of the skills and qualification of the talents. This paper aims at showcasing the possibility of matching data about skills, learning items and job offers. It opens up the opportunity to simulate career paths through visualization in order to support decision making in a world of imperfect data and information and overall this project shall also support the Europe 2020 target for inclusive growth as it aims to motivate people in acquiring new digital skills.

Selection and/or peer-review under responsibility of the organizers of ITQM 2017

Keywords: human resource management; business analytics; big data, skill matching; graph database; graph analytics; social network analysis;

1. Introduction

Gathering data about job seekers, learning items and job offers is a complicated endeavor as the quality of data is poor in all of the mentioned fields. Nevertheless, the authors and initiators of the project are strongly motivated to showcase the possibility of doing so, via using the ESCO catalogue and three different sets of algorithms, supporting the decision-making process of young talents towards their dream job.

The ESCO (European Skills, Competences, Qualifications and Occupations) classification project is supported by the European Commission with the aim to create a standard nomenclature for skills, competences,

* Corresponding author. Tel.: +43-512-2070-3500; fax: +43-512-2070-3599
E-mail address: peter.mirski@mci.edu.

qualifications and occupations and make them comparable across Europe. This is the basis for an easy communication between labor markets, employees and education institutions as well. The goal is to unify the existing semantic data from different sources like national job databases, job descriptions, assessment centers or qualification details. These data are provided in a machine-readable format using the Linked Open Data (LDO) framework [1].

The ESCO database is built on three pillars, skills and competences, occupations and qualifications. Skills and competences are the basis of the entire ESCO project, containing a collection of around 13.000 different skills/competences and around 3.000 occupations which can currently be found on the labor market. The next pillar, occupations, provides a set of occupations and occupation groups. Furthermore, these occupations are linked with skills that are required for a specific occupation. Qualifications are the formal outcome of an assessment and validation process which is obtained when a competent body determines that an individual has achieved learning outcomes to given standards.

To deploy ESCO across Europe, the data have to be provided in the 25 official EU languages. Additionally, the data structure has to be comparable with similar models like ISCO (International Standard Classification of Occupations) [2].

The project is aimed at directly linking job-seekers with Europe's industry and relevant content providers. The core idea for the future is to build an end to end solution - starting at the ESCO skill level and ending up with specific job offers. The usage and user behavior of the main stakeholders on the platform shall provide all players with anonymized, relevant and in time information about skills, learnings and jobs.

This paper provides insights in the first steps of this analytics endeavor bringing together the most important stakeholders on the labor market, namely the European Commission, industry, educational institutions and talents. This project is guided by a standard process model for data mining and analytics as they represent the basis for the further realization and implementation of the software tool which is an essential outcome of the entire OpenSKIMR project.

2. The ESCO Catalog and OpenSKIMR

OpenSKIMR is build up on the latest version of the European classification system ESCO. It can be described as a standardized terminology for skills, competences, qualifications and occupations across Europe with all its different languages. ESCO targets to bridge the gap between the world of education, training and the labor market by creating common understandable terms for skills and competences.

With OpenSKIMR a skill assessment tool using the ESCO terminology and structure should be developed. The integration of job offerings which are annotated according the ESCO descriptions allow an automated matching system between a talent's skill sets and the required skill sets of a specific job offering. Furthermore, potential skill gaps between a talent and job offering can be identified and related trainings can be proposed.

As a consequence, these requirements lead to some advanced functionalities which have to be provided by OpenSKIMR. In order to enable the matching between a talent's skill set, the ESCO occupations and real job postings, talents have to be enabled to assign the related ESCO skills in OpenSKIMR. In turn, the same must be provided for job offerings and trainings offered by educational institutions. In a first step an annotation guide to support and unify the annotation of a talent's skill set, job offerings and trainings with the ESCO nomenclature has been developed. In future releases and a stable ESCO catalog this process can be supported by the system.

ESCO currently consist of more than 14.000 skills and competences which are assigned to the ISCO hierarchy. To reduce the complexity of the assignment of the ESCO nomenclature to skill sets the number of skills has been reduced. A clustering algorithm was used to achieve this goal.

From a technical perspective, the ESCO catalog provided by the European Union is transferred to OpenSKIMR by using a classical ETL process. Besides the clustering algorithms which have to be applied the data structures have to be transformed to a graph-oriented structure as described later in this paper. The ESCO catalog is currently in pre-release status and the single datasets are still being completed.

3. Methodological Approach and Project Management

Projects in the field of big data, data science or business analytics must be guided by a comprehensive project management model. The scientific community as well as companies has developed methodological approaches for this kind of projects. The most common models are CRISP-DM (CRoss Industry Standard Process for Data Mining), SEMMA (Sample, Explore, Modify, Model, Assess), KDD (Knowledge Discovery in Databases) and TDSP (Team Data Science Process).

CRISP-DM provides a structured process model with a detailed description of the six steps with their expected outcome. CRISP-DM categorizes data mining goals and based on these categories suggestions for statistical methods are made.

SEMMA is a list of sequential steps developed by the SAS Institute which offers software tools for statistics and business intelligence. Although SEMMA can be seen as a general process model for data mining and business intelligence it also provides a logical organization of the tools provided by the SAS Institute.

KDD describes the non-trivial process to identify valid, novel, potentially useful, and ultimately understandable patterns in data. KDD consist of nine steps.

TDSP has been developed by Microsoft. The focus of TDSP is to support the collaboration of teams in their data science project and provide software tools. TDSP includes a process model which is very similar to the CRISP-DM process model but any other process model can be implemented in TDSP. It provides recommendations for managing shared analytics and storage infrastructure. TDSP provides an initial set of tools and scripts to boost the start of a data science project. TDSP is closely related to cloud based systems provided by Microsoft.

For this project CRISP-DM has been selected because it provides a data-centric methodology that is non-proprietary, application and industry-neutral, tool-neutral, and focused on business issues as well as technical issues. According to surveys CRISP-DM is still the top methodology for analytics, data mining, or data science projects [3].

In order to develop a business analytics strategy for OpenSKMR the generic phases, with the related tasks on the different abstraction levels proposed by CRISP-DM, are used as the relevant process model.

The six generic phases are described in brief as follows [4–6]:

- Phase 1 - Business Understanding: This phase focuses on understanding the project objectives and requirements from a business perspective.
- Phase 2 - Data Understanding: The data understanding phase starts with an initial data collection.
- Phase 3 - Data Preparation: This phase covers all of the activities to construct the final analytical dataset from the initial raw data.
- Phase 4 – Modeling: The modeling phase begins with the selection of the modeling approaches to be used in the specific project.
- Phase 5 – Evaluation: In this phase, the model is thoroughly assessed.
- Phase 6 – Deployment: The results and insights gained during the project have to be organized and presented in a way that the stakeholder can take advantage.

For strategy development, the most relevant phases are Business Understanding, Data Understanding and Modeling.

4. Business Understanding

The starting point is to understand the business context of OpenSKMR. The relevant stakeholders with their business goals and information needs in terms of analytics are identified.

The four main stakeholders of the OpenSKIMR pilot project are the European Commission, the industry, educational institutions and European talents. These four players have diverse goals and interests when it comes to the interaction on the job market within the European Union (EU). Nevertheless, it can be said that the overall goal of supporting talents with meaningful information in order to support their decision-making process is strongly supported by all of them.

- European Commission: The commission's interest lies mainly in gaining more insights about the interests of talents on their career plans, interesting learnings and future goals through analyzing user behavior.
- The industry: Industry shall be supported in connecting their talent needs to the public labor market through posting jobs and searching for talents.
- Educational institutions: will have the possibility to post content to a highly motivated and interested community and getting insights about relevant learnings to close the gap to the labor market.
- Talents: get matches against specific job offers and learning opportunities based on talent profiles. The possibility to simulate different routes towards specific job destinations helps to come up with better decisions.

5. Data Understanding and the OpenSKIMR Domain Model

Since data and the related data structures are the basis of all activities in analytics a profound knowledge of the available data is required. Having a clear data understanding is critical for the upcoming phases and activities.

OpenSKIMR is mainly based on the concept of social networks, the relationships between people and other entities. Implementing networks in general is a core domain of graph databases [7,8]. Highly connected objects profit from the way that the data is represented as a graph. Due to the fact that the data structure represents the real world it is easier to understand and manage the data. Another advantage of a graph database in this context is that formulating queries is much closer to a question asked in a natural language.

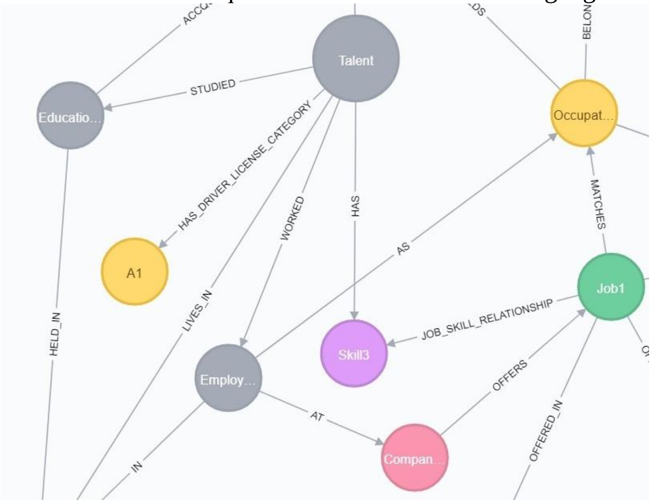

Fig. 1. Excerpt from the OpenSKIMR domain model

As a consequence, the graph database Neo4j is used for data storage. In graph databases, the domain model has to be structure in two main categories, nodes and relationships, also called graph or edge. Additionally, properties can be added to nodes and relationships.

- Nodes represent the types of domain objects used in OpenSKIMR. Each node has a label that identifies the domain object and a set of properties which store the actual values of the object. Node labels should always be nouns. Additionally, a node can relate to other nodes via relationships.
- Relationships represent connections between nodes. In contrast to node labels, the labels of relationships are usually verbs, adjectives or short phrases which describe the type of relationship.
- Like nodes, relationships may also contain properties which add further meaning to the objects and describe them further [9].

The next figure depicts a visual representation of an excerpt of nodes and relationships from the OpenSKIMR domain model. The cardinality of the relationships was intentionally left out in order to increase the readability.

The following table shows the most important nodes of the OpenSKIMR domain model:

Table 1. Nodes of the OpenSKIMR Domain Model

Node name	Node description
Talent	Personal data of the talents, including their education, known languages, skills or their formal education
Education	Formal educations, the skills gained in an educational setting, formal qualifications awarded, the organization which offers an education, EQF based rating of the education (1-8)
Employment	Employments of a talent in terms of employer (company) and the matching ESCO occupation representing the employment
Job	Job offerings by organizations (companies) with a description and the ESCO occupation matching the job, required skill for the job
Skill	Description of a sill in ESCO skill nomenclature
SkillSet	Groupings of skills
IscoGroup	Representation of the four ISCO groups and their relationships

The following table shows the most important relationships of the OpenSKIMR domain model:

Table 2. Relationships of the OpenSKIMR Domain Model

Relationships name	Relationships description
JobSkill	The required EQF level (1-8) of the skill for the job
LearningSkill:INCREASES	The EQF level granted for the skill by the learning
OccupationSkill:NEEDS	Whether or not the skill is optional or required for the occupation
TalentSkill:HAS	The skills of a talent measured in EQF levels

6. Selecting Modelling Techniques

Phase four of the CRISP-DM Methodology is the modelling phase which starts with the selection of the modelling approaches. It must be assured that the business goals defined in phase one of CRISP-DM “Business Understanding” have to be achieved by the selected modeling techniques. The business goals are defined by the different project stakeholders.

After the modeling techniques have been selected the raw data have to be transformed to a format which is required by the algorithms and the used tools. These activities belong to phase three “Data Preparation” of the CRISP-DM methodology.

As already described the data are stored as graphs in a graph database. As a consequence, the potential modeling techniques must come from the field of graph analytics, also known as network analysis.

Each graph is an ordered pair G=(V, E) consisting of a set of vertices or nodes V and a set of edges E which is an arbitrary subset of the set of all two-element subsets of V. In this context sampling algorithms can be divided into two categories, vertex (node) and edge sampling techniques.

As the term implies these methods build a subgraph G' by sampling vertices V'contained in V while leaving E unchanged, and vice versa. Random Walk (RW) [10] or Breadth-first Sampling (BFS) [11] are well established node sampling algorithms whereas Frontier Sampling (FS) [12] is a relatively new approach to edge sampling derived from RW.

Regarding performance and quality of BFS and FS the thorough work done in [13] can be used as a reference. Therefore, the authors conducted a study on these techniques based on node degree distribution and the clustering coefficient which yields a measure of how likely it is for two nodes in a graph to belong to the same cluster. They have made use of four different graph data sets including but not restricted to social networks, all provided by the Stanford Large Network Data Set Collection (https://snap.stanford.edu/data/).

The European Union as the main stakeholder of this project is interested in improving the situation of the European labor market. In order to achieve this, they need information about which skills are missing in the present and will be required in the future labor market. Hence, by fostering education of these specific skill sets early on it would be possible to bypass the risk of untrained citizens on the one hand and open job vacancies on the other hand. It is essential to gather information about particular kind of skills, e.g. those needed in the ICT sector, as well as to filter the results by country. This would enable the EU to counteract the skill shortage via targeted initiatives in respective countries.

The industry would also benefit from the information mentioned above. Furthermore, general job market trends like most popular jobs among a certain age group or country would assist companies in developing a strategic plan, adapting positions, changing job descriptions, in-house learnings and the recruiting process.

Educational institutions are mostly interested in industry requirements concerning certain skills in various fields of application in order to adapt their program according to current and future needs. The better educational opportunities are the higher the demand and the institutions reputation is. The following two approaches are of service to answering these questions.

As described in [14] network motifs can be detected which are defined as graph structures appearing in complex networks at quantities significantly higher than they would be in randomized graphs. Frequent subgraph mining can be achieved by using Fast Network Motif Detection (FANMOD) [15], for instance, or the numerically more efficient Kavosh algorithm as stated in [14]. Especially for smaller network motif sizes this one is consuming considerably less memory on average than FANMOD.

Taking a closer look at the detected network motifs which are frequent patterns of edges between talents and skill nodes, a decent prediction of which skilled employees will be available in the future should be possible. Conversely, missing edges between talents and specific fields of employment a potential deficiency can be derived. In the same way, a lack of offerings from educational institutions can be detected by looking at frequent small graph structures inside the entire network regarding education and the skills they improve.

In contrast to rather localized interactions occurring, for example between fellow talents or coworkers, it is also interested to examine the graph structure on a large scale, e.g. nationwide. An arbitrary graph can be defined as follows: An arbitrary graph H is called a topological minor of G if a subdivision of H, i.e. essentially the same graph but with more nodes on existing edges, is contained in G as a subgraph. In [16] the authors propose the TSMiner algorithm which mines frequent constrained topological patterns that are present as subgraphs in some given network. Considering the OpenSKIMR data set, this algorithm is of particular interest if the entire graph

G with all its individual occupations, jobs and users contains a subgraph H with less vertices and edges but increased visibility of relations between companies.

Generally speaking, the TSMiner algorithm is able to fulfill the demands of European Union, industry and users in the same way as FANMOD or Kavosh do but on a larger scale. Thus, if future job trends of some country or even EU-wide are of interest this algorithm yields frequent large-scale structures of edges and nodes accordingly.

A typical user of the OpenSKIMR platform targets receiving essential information for his/her personal career path like skills currently in high demand and in which country job offers are available. After gathering this information, he/she can work out an educational path and figure out where to reside. A lack of skills needed for an employment can be rectified by attending the right learning or possibly learnings. In any case one wants to find the shortest learning route from status quo to job offer. That is what is needed for the following algorithm.

A very well-known shortest path algorithm was introduced by Dutch computer scientist Dijkstra in 1959 [17]. It has a time complexity of $O(n^2)$ and was later optimized with respect to its runtime by Fredman [18] to $O(n\log n)$ which still is less than perfect. Especially when considering real time user interaction with the OpenSKIMR platform this is impractical. In 2016 a novel method with increased performance was introduced in [19].

The shortest path algorithm based on community detection (SPCD) utilizes the fact that communities in a graph consist of vertices sharing common properties hence having denser connections. The authors tested their algorithm on five different data sets and evaluated its improved performance relative to Dijkstra’s algorithm. On the average, SPCD ran about 200 times faster than its predecessor. With the aid of SPCD OpenSKIMR is able to specify the education (represented as nodes in the graph) a user has to complete to be properly trained for a specific job.

In order to detect communities in social networks as mentioned above an algorithm that focuses on not only efficiency but also completeness is needed. State-of-the-art techniques for finding maximal cliques divide the graph into smaller sections and search for cliques in each one of them separately. However, there is always a trade-off between computation time and how severe the error one introduces turns out to be. According to [20] this kind of adjustment is not a necessity because their approach is able to meet both demands. Provided that the underlying graph is sparse, which certainly is the case with OpenSKIMR, the Find-Max-Cliques approach relies on a two-way partitioning of the network paired with a standard maximal clique enumeration algorithm, e.g. Bron-Kerbosch Pivot (BKPivot) [21].

7. Future Outlook and Further Implementation

This first step of the analytics activities follows a business oriented approach to identify algorithms and techniques that can be used to fulfill the business goals of the project’s stakeholders. Once the software tool which is part of the OpenSKIMR project is deployed real-life data are available. The selected analytics models have to be tested and evaluated.

Following the CRISP-DM methodology the next steps of the project can be found in phase four, the modeling phase. The next steps are a) defining the tests for the selected modeling techniques, b) building and implementing the models and c) assessing the models.

Acknowledgements

This paper and the underpinning research work was enabled through a funding of the European Commission, the Directorate-General for Communications Networks, Content and Technology.

References

[1] Smedt JD, Le Vrang M, Papantoniou A. ESCO: Towards a Semantic Web for the European Labor Market.

[2] European Commission. ESCO, European classification of skills: The first public release. Luxembourg: Publications Office of the European Union; 2013.

[3] Piatetsky G. Top methodologies for Analytics. [July 25, 2017]; Available from: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

[4] Abbott D. Applied predictive analytics: Principles and techniques for the professional data analyst. Indianapolis, Ind.: Wiley; 2014.

[5] Lin N. Applied business analytics: Integrating business process, big data, and advanced analytics. Upper Saddle River, NJ: Pearson; 2015.

[6] Putler DS, Krider RE. Customer and business analytics: Applied data mining for business decision making using R. Boca Raton FL u.a.: CRC Press; 2012.

[7] Otte E, Rousseau R. Social network analysis: A powerful strategy, also for the information sciences. Journal of Information Science 2016;28(6):441–53.

[8] Angles R, Prat-Pérez A, Dominguez-Sal D, Larriba-Pey J. Benchmarking database systems for social network applications. In: Boncz P, Neumann T, editors. First International Workshop on Graph Data Management Experiences and Systems - GRADES '13. New York, New York, USA: ACM Press; 2013, p. 1–7.

[9] Daly EM, Haahr M. Social network analysis for routing in disconnected delay-tolerant MANETs. In: Kranakis E, Belding E, Modiano E, editors. Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing - MobiHoc '07. New York, New York, USA: ACM Press; 2007, p. 32.

[10] Gjoka M, Kurant M, Butts CT, Markopoulou A. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In: 2010 Proceedings IEEE INFOCOM. IEEE; 2010, p. 1–9.

[11] Wilson C, Boe B, Sala A, Puttaswamy KP, Zhao BY. User interactions in social networks and their implications. In: Schröder-Preikschat W, Wilkes J, Isaacs R, editors. Proceedings of the fourth ACM european conference on Computer systems - EuroSys '09. New York, New York, USA: ACM Press; 2009, p. 205–218.

[12] Ribeiro B, Towsley D. Estimating and sampling graphs with multidimensional random walks. In: Allman M, editor. Proceedings of the 10th annual conference on Internet measurement - IMC '10. New York, New York, USA: ACM Press; 2010, p. 390–403.

[13] Wang T, Chen Y, Zhang Z, Xu T, Jin L, Hui P et al. Understanding Graph Sampling Algorithms for Social Network Analysis. In: 31st International Conference on Distributed Computing Systems Workshops. IEEE; 2011, p. 123–128.

[14] Kashani ZRM, Ahrabian H, Elahi E, Nowzari-Dalini A, Ansari ES, Asadi S et al. Kavosh: A new algorithm for finding network motifs. BMC Bioinformatics 2009;10.

[15] Wernicke S, Rasche F. FANMOD: A tool for fast network motif detection. Bioinformatics 2006;22(9):1152–3.

[16] Jin R, Wang C, Polshakov D, Parthasarathy S, Agrawal G. Discovering frequent topological structures from graph datasets. In: Grossman R, Bayardo R, Bennett K, editors. Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05. New York, New York, USA: ACM Press; 2005, p. 606–611.

[17] Dijkstra EW. A note on two problems in connexion with graphs. Numer. Math. 1959;1(1):269–71.

[18] Fredman ML, Tarjan RE. Fibonacci Heaps And Their Uses In Improved Network Optimization Algorithms. In: 25th Annual Symposium onFoundations of Computer Science, 1984. IEEE; 1984, p. 338–346.

[19] Gong M, Li G, Wang Z, Ma L, Tian D. An efficient shortest path approach for social networks based on community structure. CAAI Transactions on Intelligence Technology 2016;1(1):114–23.

[20] Conte A, Virgilio RD, Maccioni A, Patrignani M, Torlone R. Finding All Maximal Cliques in Very Large Social Networks Technology, EDBT 2016, Bordeaux, France, March 15-16, 2016, Bordeaux, France, March 15-16, 2016, p. 173–184.

[21] Bron C, Kerbosch J. Algorithm 457: Finding all cliques of an undirected graph. Commun. ACM 1973;16(9):575–7.