Comparing Competences on Academia and Occupational Contexts based on Similarity Measures

Alexandra González Eras , Pablo Quezada S. , Patricia Ludeña González and Carolina Gallardo

Keywords: Body of Knowledge, Machine Learning, Similarity Measures, Professional Profile, DISCO II, VSM, NLP.

Abstract: This paper is a first contribution of a schema to match professional and work competences through similarity measures. In this contribution we focus on the determination of connections between university profiles based in standards (body of knowledge and thesauri) similarity measures and Natural Language Processing (NLP) techniques. Our first experiments proved that this hybrid schema got a promise results in the recognition of competency patterns in order to apply in the laboral context.

1 INTRODUCTION

One of the main concerns of the software industry is to develop the talent of its human resources, since; the quality and innovation of its products and services depend on a great extent of knowledge, and the ability and the talent of software engineers. Therefore, the relationship university-employment becomes a cycle comparison of skills, college profiles which meet the competencies of future professional while labor profiles have the skills required to fill a job position.

However, in reality it is almost impossible to compare competencies, mainly due a problems as: incompatible profiles (Fazel-Zarandi and Fox, 2013) (Stevens, 2013) and unstructured, ambiguous, and sometimes incomplete data (Fazel-Zarandi, 2013). Looking for a solution, models and platforms have been proposed in order to profile standardization (Draganidis and Mentzas, 2006) and comparison through competency frameworks (e-CF[1] , SIOC[2], O*NET[3]). Nevertheless, the actors rarely use these tools, or they have only been proposed for one language context without a real application in others. As a result, is difficult to obtain a standardization of profiles that permit the comparison of competence.

[1]e-Qualifications Framework, available online at http://www.ecompetences.eu

[2]Semantically-Interlinked Online Communities, available online at http://sioc-project.org

[3]Occupational Information Network, available online at https://www.onetonline.org

This paper is a first approximation to develop a strategy in order to compare professional and work competences. In order to develop the first experiment, we used the standard DISCO II to compared with college profiles and thereby obtain a middle ground that allows us to meet the reporting inconsistencies. In the same context, we focus on the combination of similarity measures (Harispe et al., 2013), (Turney et al., 2010), (Turney, 2006) and NLP techniques based on n-grams to find common patterns between university profiles and DISCO II.

2 CONTEXT

The Fig. 1 shows a picture about the context of this research where university profiles are showed as circles, job offers as triangles and competencies are represented by the colors of the figures.

We realize that competencies have a different degree of presence in the profiles, as represented by the different sizes of the figures. Also, there are groups of profiles (academic or curriculums), covering a greater or lesser extent job offers; also they share competencies in each context, besides new skills required in the job offers. Otherwise, standards offer a competence computational representation, that sometimes it is insufficient to show competency meanings.

Therefore, we comprise the main issues of this research in the following main points:

1. Job market always requires competencies that

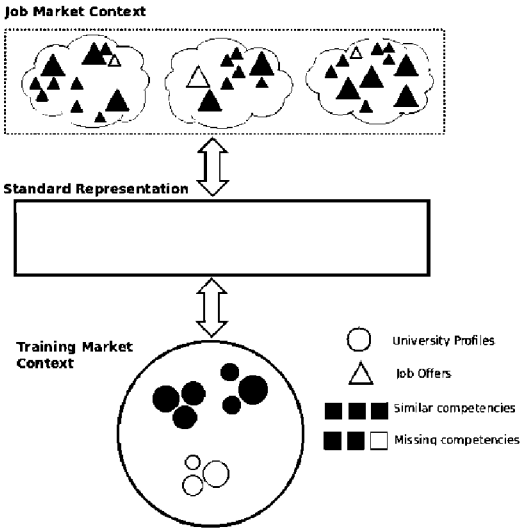


Figure 1: Context problematic.

Universities have problems to cover (new/missing competencies). Sometimes the profiles have different structure and lack of information. Besides, the profiles clearly not describe the competencies or the competency elements, so we cannot make a comparison between them (Paquette et al., 2012). The job profiles are more close to describe job features like activities or roles and not like competencies (Malzahn et al., 2013). Moreover, the curriculum for a university study normally gives only dependencies between courses and a generic description de capacidades en un area especfica (Dorn and Pichlmair, 2007). These descriptions do not have a clear relation with competencies and por ende es difcil identificar nuevas competencias as como a aquellas competencias que fueron omitidas o no cubiertas por el perfil (Fazel-Zarandi, 2013).

2. Job offers and university profiles share competencies, but with different level of meaning (different conceptualization). Mainly due to different interpretations that each actor has about the definition of competency (Fazel-Zarandi and Fox, 2009). For instance, the term competency appears to be used at times to refer to actions and their consequences and at others to refer to cognitive skills and personality characteristics (Stevens, 2013) , or the competencies of individuals may be expressed in terms of qualifications and certifications, such as academic degrees (Malzahn et al., 2013) or as learning outcomes within educational processes (Paquette, 2007). Other aspect to consider is the degree of performance of competences in which the labor market related activities and functions (Bizer et al., 2005), whereas in academia is related to generic descriptions of occupational areas (Dorn and Pichlmair, 2007).
3. Standards have few flexible and complex schemas. Standards have the function of facilitate the exchange of competence descriptions. Standards as : IMS RCDEO 2002, IEEE RCD 2004, HR-XML are the core of many Applications of Manager of Competences (Draganidis and Mentzas, 2006). In the same context each individual prefers to use a free vocabulary to convey their competencies (Fazel-Zarandi and Fox, 2009). Similarly thesauri and taxonomies have been proposed to define skills and competencies (SIOC, O*NET, DISCO II) but, stakeholders do not use these schemas to create their documents (Malzahn et al., 2013), because they often lack motivation to add them into their profiles (Hansen et al., 2011) because the specifications are difficult to apply (Malinowski et al., 2006), therefore over time the profiles become out-dated (Fazel-Zarandi and Fox, 2013).

3 RELATED WORKS

3.1 NLP Techniques

NLP techniques contribute to the recognition of the text portions used for comparison. In this sense, many of works in the literature focused on the connection of ontologies With NLP tools, to extract competencies as Entities and relations. In (Malzahn et al., 2013) The NLP techniques are used to extract patterns Competency. In (Janev and Vranes, 2011) a framework that extracts patterns of personal and enterprise skills since different text sources used. (Yahiaoui et al., 2006) uses a semantic annotation scheme based on a multi ontology framework for labeling CVs and job offers by STI concepts instances.

3.2 Similarity Measures

When we think about similarity, we associate the perspective of two entities sharing in some degree a set of characteristics, and a similarity measure capturing the strength of this semantic interaction in connection with their meaning (Turney et al., 2010). Measures could estimate the similarity/dissimilarity between specific kind of semantic representation on which is based the comparison, for instance units of language as words, sentences, paragraphs and documents (Harispe et al., 2013).

This idea represent the semantic between Vector Space Model (VSM) in which Vector Space Model

(VSM), is consider how as a point in a space (a vector in a vector space). Points that are close together in this space are semantically similar and points that are far apart are semantically far (Turney et al., 2010). For instance, to measure the similarity between a query and a document (Manning et al., 2008), or within algorithms that measures the similarity of semantic relations (Turney, 2006).

VSMs have several attractive features. VSMs extract knowledge automatically from a given corpus, thus they require much less Labour than other approaches to semantics, such as lexicons, thesauri and ontologies (Manning et al., 2008).

In the same context, VSM is used in job recruitment recommender systems para rankear profiles based in candidate-offer matrixs (Linden et al., 2003), and the hybrid schema where with the combination of NLP techniques and similarity measure for example Coseno, to cluster the competences (Malzahn et al., 2013). In (Buitelaar and Eigner, 2008) linguistic patterns of knowledge are combinated with distance measures like Euclidean L1norm to build semantic networks based on term frequency. In (Reichhold et al.,) a measure of role relevance added to Cosine distance match job offer vectors.

4 METHODOLOGY

In order to develop this paper we proposed the next methodology based in 3 stages. (Figure 2).

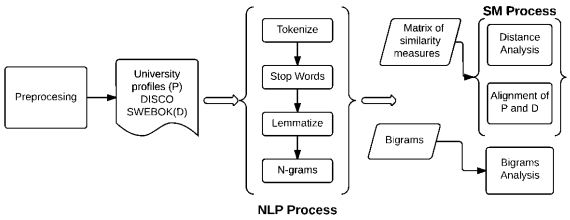


Figure 2: First proposal of a hybrid schema to get a baseline.

4.1 Preprocessing

The sources for our corpus was taken from the following sources:

- Career profiles of Software Engineering and related, which were taken from university websites of Latin American universities, selected from the database of Webometrics[4].

[4]Ranking Web of Universities, available at http://www.webometrics.info/

In order to choose the candidates phrases in university profiles we select paragrahps of the sections: description, occupational field, skills, competences and knowledge areas, in which show competences sentences. In the Figure 3 we can show these competences.

Example 1: Demostrar conocimientos de algoritmica y programación
Example 2: Analizar, diseñar e implementar sistemas de bases de datos gerenciales

Figure 3: Example of academic competencies.

To resolve the problem of complexity of sentences we divided the paragraphs following the proposed scheme: abilities represent in verbal structure and some topics related with the knowledge area that represent the nominal structure. In the case of complex sentences (more than one verb) we perform sentence repetition based on the following rule: for each verb in a sentence, a new sentence will be added which is the same sentence but with one verb.

- Standards: we are considering the standard DISCO II which offers competence phrases that have been chosen based on a consensual process.

4.2 NLP Process

The process of NLP is based on a superficial representation of texts (Metzler et al., 2007), in which we submit the corpus to the following tasks:

1. Tokenization, stop words: words are separated according to the spaces between them based on the Snowball list of stop words for the Spanish language, then a manual review of the words, to eliminate cases that were not considered (articles, prepositions, conjunctions and numbers).
2. Lemmatization: to regularize surface variations of words by converting them to the same form with reducer el sparsity problem. The types of normalization applied were: case folding (converting all words to lower case) and lemmatization (changing the words by its canonical form).
3. N-grams: we obtained unigrams and bigrams since university profiles (P) and standard (D), with the propose of develop a first identification of common pattern between corpus.

4.3 Similarity Measures Process (SM Process)

In the SM process we focused on the combination of similarity measures as example Cosine 1, which al-

lows comparing two frequency vectors whether they be raw o weighted or they have different lengths (Turney et al., 2010). Cosine captures the idea that the length of the vectors is irrelevant; the important thing is the angle between the vectors (Deerwester et al., 1990).

$$\cos(x,y) = \frac{\sum_{i=1}^{n}(x_i * y_i)}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \quad (1)$$

As a complement, we used a lexical character 2, that calculated the relation between unigrams and bigrams, giving high weight to similar bigrams. Then, we built a matrix of similarities between both corpuses (D x P matrix) counting the number of similar unigrams (nu) and the number of similar bigrams (nb). The value of similarity gave more importance to bigrams. In the same context we develop a distance analysis where D x P matrix was submitted to VSM, which included Singular Value Decomposition technique to reduce the Sparsity. Additional we develop a n-grams in order to search patters compose by verbal and nominal structures.

$$SM = nu * 0.5 + nb \quad (2)$$

5 EXPERIMENTATION

With the goal of matching university profiles (P) with the standard DISCO II (D), we conducted the following experiments:

- Distance Analysis: to compare both corpus and get the distances between them. We also use the entropy to determine whether it is possible to compare P and D with no other additional information.
- Perfect alignment of P and D: to establish the similarity between P and D on the basis of different types of alignment.
- Bigrams analysis: to detect the occurrence of patterns.

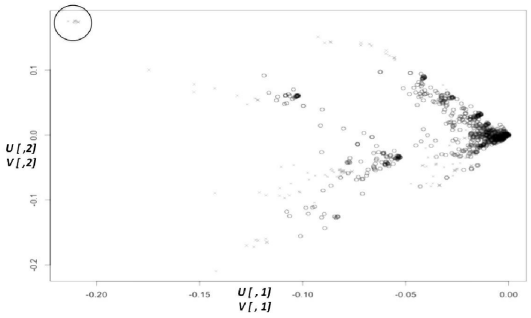
5.1 Distance Analysis

For the analysis of distances, D x P matrix was then submitted to SVD under the framework of a distance matrix and cosine as a distance measure. The Table1 shows preliminary results:

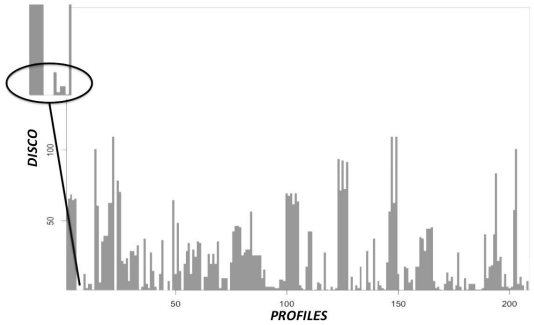
We get that the maximal distance is 1.33, the average distance is around 0.82, and the standard deviation is around 0.21. With these numbers, we clearly know that most of the sentences are dissimilar. The Figure 4 shows the results of the SVD in 2D.

Table 1: Results of Distance Analysis between P and D.

Criteria	Values
Maximal distance	1.3271
Average distance	0.8212
Standard deviation	0.2082



(a) SVD of D versus P (SVD = U * S * V^t : where U= profiles and V= disco)



(b) Disco versus Profiles (DxP)

Figure 4: Plot 2D of SVD over D and P.

In the second plot we can confirm the dissimilarity between D and P (the white space), but also we notice that there is a lot of redundancy in P since many sentences differ from others by only one token. This is due to the fact that P contains many repetitions To the diagonal analysis we start with a following hypothesis: the diagonal of the distance matrix is distorted when information is missing. The Figure 4 confirms that D and P have a distorted diagonal, which means that D and P share few similar sentences. We can see these sentences in strong red. We can see peaks in the graph, which means that some sentences in P are similar to many sentences in D. Besides; many sentences of P (the white area over the phrase 100) are not covered at all in D (in x).

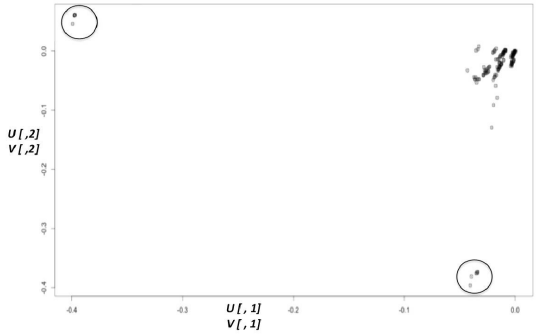
5.2 Alignment of Corpus

To achieve a measure of similarity, the purpose is to

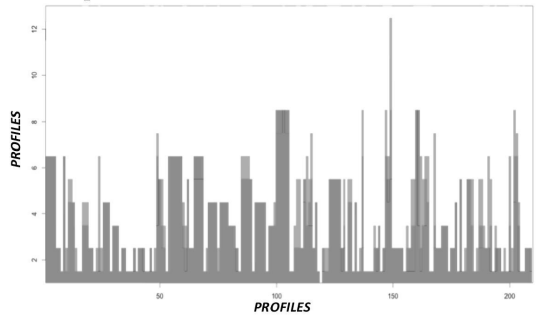
find a perfect alignment between D and P. Then, How does it looks when the alignment is perfect that is if the profiles are just like the standard?. Likewise, we made some tests with the purpose of compare D and P.

5.2.1 Comparing P with P

The Figure 5 shows plots of SVD over P with P, which we can see that the similarity matrix is just perfect. Also, the matrix of cosine distance gives a good diagonal.



(a) SVD of P versus P (SVD = U * S * V^t : where U and V= profiles)



(b) Profiles versus Profiles (PxP)

Figure 5: Plots 2D to compare of P with P.

Although, some sentences in P are redundant (big rectangles in the diagonal). So, some sentences in P are similar with many others (up to 6 or 7).

5.2.2 Comparing D with D

The Figure 6 shows the plots of SVD between D with D, we can see that the similarity matrix of D with D.

The matrix of cosine distance gives reveal of course a diagonal, but not so clean. It means the D is not so clean as we thought. Many sentences are similar to others; this is due to the fact that the entire hierarchy of DISCO II is used. Besides, the plot shows many sentences of D are similar with others in D (as shown by the peaks).

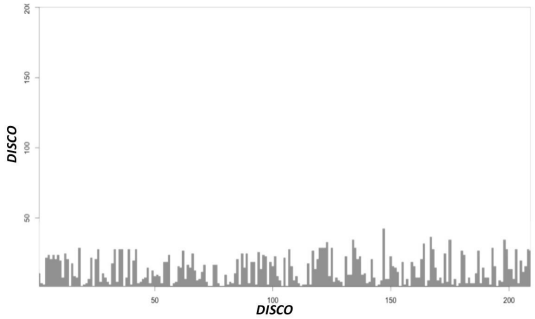


Figure 6: Plots 2D to compare of D with D.

Count	Bigram
182	('sistema', 'informático')
120	('sistema', 'información')
60	('base', 'dato')
44	('sistema', 'operativo')
28	('programa', 'software')
24	('seguridad', 'informático')
16	('administrar', 'sistema')
15	('resolver', 'problema')
10	('solución', 'informático')
9	('hardware', 'software')
9	('sistema', 'seguridad')
8	('diseñar', 'sistema')
8	('utilizar', 'herramienta')
8	('diseño', 'sistema')
6	('aplicación', 'informático')

Count	Bigram
3	('diseñar', 'red')
3	('seguridad', 'dato')
3	('software', 'hardware')
2	('proceso', 'desarrollo')
2	('realizar', 'función')
2	('diseñar', 'implementar')
2	('diseñar', 'solución')
2	('implementar', 'sistema')
2	('administrar', 'red')
2	('proceso', 'negocio')
2	('programar', 'aplicación')
2	('identificar', 'problema')
2	('gestionar', 'proyecto')
2	('solución', 'problema')
2	('problema', 'hardware')

Figure 7: First proposal of a hybrid schema to get a baseline.

In 7, we can show in the remarked bigrams which involve a verb like ”gestionar” or a nominalization like ”diseo” followed by a noun, these kind of bigrams very often express competence. Therefore, we suppose that removing the verb we will remove many competences.

Besides we found bigrams like base dato and sistema which gives us a idea about common patterns and the posibles patterns for new experiments, as the case of mapping of domain areas.

6 ANALYSIS AND DISCUSSION

The comparison scheme based on unigrams and bigrams and similarity measures is promising to find similar items. The results obtained in the bigrams analysis confirm the need to explore other possibilities of combinations between noun phrases and verb phrases. The use of patterns can provide new possibilities for the interpretation of isolated values.

There is much redundancy in P and D and it is necessary to perform a pre-cleaning over the corpus before applying any comparison scheme. The interpretation of negative values (isolated) is not clear in the case of P x P matches. We cannot be certain that

these values are equivalent to missing information or a derivative problem of redundancy in P and D. In future experiments should address the analysis of the specific statements that cause these values.

The comparison of P and D could be performed at different levels of the standards hierarchy, in order to reduce the number of outliers. To achieve greater entropy in the matching we have to consider the use of parts of DISCO II rather than all the entire hierarchy and thus reach a lower level of redundancy in P x D.

7 CONCLUSIONS AND FUTURE WORKS

We propose a model to compare university and job offer profiles based on similarity measures.

The use and combination of different similarity measures to get a high performance will be developing. Also, using standards to framework construction and validation permit give a solution of vocabulary mismatch problem.

As a result, we expect to get similarity indicators between university and job offer profiles.

We propose a comparison with job profiles based on competencies and guides the referents in engineering context.

ACKNOWLEDGEMENTS

We thank our colleagues from Politechnical University of Madrid and Universidad Técnica Particular de Loja who provided insight and expertise that greatly assisted the research of this paper.

REFERENCES

Bizer, C., Heese, R., Mochol, M., Oldakowski, R., Tolksdorf, R., and Eckstein, R. (2005). The impact of semantic web technologies on job recruitment processes. In *Wirtschaftsinformatik 2005*, pages 1367–1381. Springer.

Buitelaar, P. and Eigner, T. (2008). Topic extraction from scientific literature for competency management. In *The 7th International Semantic Web Conference*.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407.

Dorn, J. and Pichlmair, M. (2007). A competence management system for universities. In *ECIS*, pages 759–770.

Draganidis, F. and Mentzas, G. (2006). Competency based management: a review of systems and approaches. *Information Management & Computer Security*, 14(1):51–64.

Fazel-Zarandi, M. (2013). *Representing and Reasoning about Skills and Competencies over Time*. PhD thesis, University of Toronto.

Fazel-Zarandi, M. and Fox, M. S. (2009). Semantic matchmaking for job recruitment: an ontology-based hybrid approach. In *Proceedings of the 8th International Semantic Web Conference*.

Fazel-Zarandi, M. and Fox, M. S. (2013). Inferring and validating skills and competencies over time. *Applied Ontology*, 8(3):131–177.

Hansen, D. L., KHOPLAR, H., and Zhang, J. (2011). Recommender systems and expert locators. *Understanding Information Retrieval Systems: Management, Types, and Standards*, pages 435–447.

Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2013). Semantic measures for the comparison of units of language, concepts or instances from text and knowledge base analysis. *arXiv preprint arXiv:1310.1285*.

Janev, V. and Vranes, S. (2011). Ontology-based competency management: the case study of the mihajlo pupin institute. *J. UCS*, 17(7):1089–1108.

Linden, G., Smith, B., and York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80.

Malinowski, J., Keim, T., Wendt, O., and Weitzel, T. (2006). Matching people and jobs: A bilateral recommendation approach. In *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, volume 6, pages 137c–137c. IEEE.

Malzahn, N., Ziebarth, S., and Hoppe, H. U. (2013). Semi-automatic creation and exploitation of competence ontologies for trend aware profiling, matching and planning. *Knowledge Management & E-Learning: An International Journal (KM&EL)*, 5(1):84–103.

Manning, C., Raghavan, P., and Schütze, H. (2008). Introduction to information retrieval. cambridge university press, cambridge, uk.

Metzler, D., Dumais, S., and Meek, C. (2007). *Similarity measures for short segments of text*. Springer.

Paquette, G. (2007). An ontology and a software framework for competency modeling and management. *Educational Technology & Society*, 10(3):1–21.

Paquette, G., Rogozan, D., and Marino, O. (2012). Competency comparison relations for recommendation in technology enhanced learning scenarios. In *Proceedings of the 2nd Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2012)*, volume 896, pages 23–34. Citeseer.

Reichhold, M., Kerschbaumer, J., Fliedl, G., and Winkler, C. Automatic generation of user role profiles for optimizing enterprise search.

Stevens, G. W. (2013). A critical review of the science and practice of competency modeling. *Human Resource Development Review*, 12(1):86–107.

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Turney, P. D., Pantel, P., et al. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Yahiaoui, L., Boufaïda, Z., and Prié, Y. (2006). Semantic annotation of documents applied to e-recruitment. In *SWAP*.