International Journal of Artificial Intelligence in Education (2019) 29:536–567
https://doi.org/10.1007/s40593-019-00185-z

ARTICLE

Determination of Professional Competencies Using an Alignment Algorithm of Academic Profiles and Job Advertisements, Based on Competence Thesauri and Similarity Measures

Alexandra González-Eras[1,2] · Jose Aguilar[2]

Published online: 3 September 2019

Abstract

Describing the competencies required by a profession is essential for aligning online profiles of job seekers and job advertisements. Comparing the competencies described within each context has typically not be done, which has generated a complete disconnect in language between them. This work presents an approach for the alignment of online profiles and job advertisements, according to knowledge and skills, using measures of lexical, syntactic and taxonomic similarity. In addition, we use a ranking that allows the alignment of the profiles to the topics of a thesaurus that define competencies. The results are promising, because the combination of the measures of similarity with the alignment with thesauri of competencies offers robustness to the process of generation of professional competence descriptions. This combination allows dealing with the common problems of synonymy, homonymy, hypernymy/hyponymy and meronymy of the terms in Spanish. This research uses natural language processing to offer a novel approach for assessing the match of the competencies described by the applicants and by the employers, even if they use different terminology. The resulting approach, while developed in Spanish for computer science jobs, can be extended to other languages and domains, such is the case of recruitment, where it will contribute to the creation of better tools that give feedback to job seekers about how to best align their competencies with job opportunities.

Keywords Professional competencies · Academic profiles · Job advertisements · Competence thesauri · Similarity measures · Alignment of profiles

✉ Alexandra González-Eras
acgonzalez@utpl.edu.ec

Jose Aguilar
aguilar@ula.ve

[1] Departamento de Ciencias de la Computación y Electrónica, Universidad Técnica Particular de Loja, San Cayetano Alto, Loja, Ecuador

[2] CEMISID, Facultad de Ingenieria, Universidad de Los Andes, Mérida, Venezuela

Introduction

Throughout the last decades, the concept of competence has gained relevance, not only in the workplace (Smirnov et al. 2016), but also in the academic field (Fazel-Zarandi 2013; Paquette 2007), where the knowledge of the competencies that are required for a profession is of great importance for the update of professional profiles (e.g., job advertisements) and curricula (e.g., academic profiles) (Paquette 2016). In this paper, the "academic profile" term meaning is the academic competencies (formation), and in a general way, is the online profiles of job seekers, students, academic websites, etc. In particular, the comparison between academic profiles and job advertisements makes possible to identify which ones have similar competencies, and which competencies should be added so that an academic profile, e.g., a student's LinkedIn page, an online resume, or a graduate student website, can be matched to a job opportunity in a job listing website, like Monster and Indeed websites.

On the other hand, the developments in Web technologies and AI techniques to build the Semantic Web have allowed a new set of applications with important implications for Web-based education (Aguilar et al. 2015). One possible utilization consists in the characterization of the competencies based on the professional profiles and curricula. However, this comparison presents difficulties, mainly due to the way in which competencies are expressed in both contexts (Paquette et al. 2012). For example, in the academic context they are manifested as learning outcomes (Worsley and Blikstein 2018), while in the work context the competencies are presented as functions, knowledge areas, or skill levels in specific subjects (Rácz et al. 2018; Rosa et al. 2015). Consequently, there is a problem of understanding the meaning of competencies, such that one competence can be similar to another, even though the same words are not used to express them.

Based on the definition that competence is something "that can demonstrate the application of a generic skill on some knowledge" (Paquette et al. 2012; Paquette 2007), the purpose of the present work is the development of a comparison scheme between competencies, which allows to overcome the problem of ambiguity among them, using similarity measures of texts, combined with thesauri. To do this, two lexical measures are used, such as Levenshtein and Dice's, to determine the levels of coincidence between knowledge and skills topics of the academic profiles and job advertisements. Then, according to a threshold, those with the greatest lexical similarity are chosen. Then, we use the taxonomic structure of the thesaurus to obtain a measure of semantic similarity of knowledge and skill topics, inspired by the Ant Colony Optimization algorithm. First, the levels of coincidence of the topics in the thesaurus are identified, to later determine the highest similarity through the analysis of ancestors, brothers and sons of each one of these topics (González-Eras and Aguilar 2015; Mendonza et al. 2015; González-Eras et al. 2017; Guevara et al. 2017). As a result, a similarity is obtained from the analysis of competencies of each profile, which considers not only topic characteristics, but also their context.

There are research efforts in which semantic representations and similarity measures are used to compare processes or models, and thus solve ambiguity problems in textual expressions, caused by the use of synonyms, homonyms, or different levels of abstraction in the description of entities or concepts. In (Ehrig et al. 2007) it is obtained the similarity between business process models, represented in a Petri network, with two

measures, the first establishes the lexical distance between pairs of concepts and a dictionary, to determine synonyms, and the second, is a structural measure that recognizes homonyms between concepts, comparing their position in each model. Likewise, in (Dijkman et al. 2011) the similarity between the processes is through the edit distance between process names, and then a weighting of relationships intersection of common names and synonyms of names not common in chains is performed. In (Van Dongen et al. 2013) similarity measures are used to compare business process models: measures of similarity of process names, which measure the similarity between words and structural similarity measures, which in addition to aligning process names, also measure the relationships between them. As for the comparison of competencies, in the work of (Malzahn et al. 2013) similarity measures are used to compare entities and competencies in professional profiles, first according to their editing distance (Levenshtein), then with the support of thesauri (Germa-Net) and dictionaries (Wortschatz), to detect synonyms, and, through semantic measures, to align concepts according to their frequency.

For the present work, we propose the implementation of similarity algorithms that make an alignment of the knowledge and skill topics found in academic profiles and job advertisements, against the topics present in a competence thesaurus. Firstly, by means of a lexical measure, that compares them letter by letter, and once the topic of the thesaurus of greater similarity is found, it uses a measure of structural similarity to verify that they have equal ancestors, brothers and sons within the thesaurus taxonomy (González-Eras and Aguilar 2015). This establishes a semantic measure for the competence alignment between academic profiles and job advertisements, based on the similarity of their knowledge and skill topics. If our approach works well, the metrics will indicate the curricular topics with which the job listings are most aligned.

This article is structured as follows: first, the characterization of competencies is carried out in the context of academic profiles and job advertisements, followed by the similarity in the case of competencies; then the architecture of the proposal is addressed, and then the experimentation, the analysis of results, and the conclusions of this work.

Characterization of the Competence Concept

The context that will be used during the article, to explain the different concepts used in it, is the domain of Computer Science. The objective is to make the comparison of academic profiles and job advertisements according to the competencies; For this, functional or specific competencies are analyzed, based on the concept that says that a competence is defined by "the ability with which a professional develops in a specific area of knowledge" (Paquette et al. 2012). Thus, we understand as competence the constituent elements of skill and knowledge, since knowledge includes the set of topics or issues that are part of a profession that are necessary to function in it (De Leenheer et al. 2010), while skill represents the capacity to use knowledge to act successfully in the development of an activity (Beckers 2011; Blanco-González et al. 2011).

In common practice, the competencies representation has been carried out through linguistic declarations, which do not formally describe the domains of knowledge or

skill, in addition to not being suitable for computational processes; which makes it difficult to compare competencies in job advertisements and academic profiles. In addition, for each type of profile, the sentence structure containing the competencies is different. Table 1 shows examples of the text structures found in the profiles. As we can see, the expressions highlighted in red represent skills, while the expressions highlighted in blue represent knowledge.

Although these statements demonstrate the presence of skills and knowledge, sentences have different lengths, present more than one verb to denote skill levels, and use different words to express the same knowledge. Consequently, comparing profiles based on these statements implies the alignment of knowledge topics, to establish similarities between ambiguous topics; and the alignment of skill topics, in order to select those skills that represent the competence.

In general, the ambiguities that can be found in knowledge topics correspond to synonymic relationships, where two topics have the same meaning, although they are written differently. For example, the topics "parallel processing" and "distributed computing" are similar terms since they share the same knowledge area. There are also cases of hyponymy / hypernymy, where a topic has a hierarchical semantic relationship with another topic, for example, "systems" with "operating systems" or "distributed systems"; and the relations of meronymy where the topics share the same hierarchical level, as is the case of "programming languages" with "Java language" and "PHP language" (Lundqvist et al. 2011). In the same way, the topics of skill: "demonstrate", "indicate" and "expose" share a semantic relationship because they are, according to dictionaries and thesauri, synonyms (Ortiz Sánchez 2016).

Table 1 Examples of competences found in profiles

Competencies of labor profile	Competencies of academic profile
Diseñar arquitecturas basadas en computación en paralelo. (Design architectures based on parallel computing)	Conocimientos de software basado en computación distribuida, montaje y utilización de redes de interconexión entre equipos de cómputo. (Knowledge of software based on distributed computing, assembly and use of interconnection networks between computing equipment.)
Interactuar con bases de datos y lenguaje SQL. (Interact with databases and SQL language.)	Conocimiento de sistemas operativos Linux y Windows, dispositivos periféricos y equipos electrónicos involucrados en el control de procesos industriales. (Knowledge of Linux and Windows operating systems, peripheral devices and electronic equipment involved in the control of industrial processes.)
Abordar proyectos de sistemas informáticos. (Address computer systems projects.)	Diseñar y administrar sistemas de comunicación de datos, además de la toma de decisiones y en la difusión de las mejores opciones de desarrollo de software. (Design and manage data communication systems, as well as making decisions and disseminating the best software development options.)

Semantic Sources

One way to resolve cases of textual ambiguity is to align text units against semantic structures, thus obtaining their similarity (Harispe et al. 2013). For competencies, the semantic sources normally used are thesauri, taxonomies and dictionaries in the same language and domain of knowledge. For this reason, two of these semantic structures were used for our research: the DISCO II thesaurus[1] and a thesaurus based on the BLOOM taxonomy proposed in (Ortiz Sánchez 2016).

The DISCO II thesaurus is an international standard used in the creation of competence profiles in the labor and educational fields, which has a Spanish version. Furthermore, the Computer Science area includes statements that paraphrase competencies, which contain knowledge elements that represent learning outcomes (Müller-Riedlhuber 2009). It is a controlled vocabulary, and the existing relationships between terms are of three types: 1. Semantic equivalences, which represent synonyms, 2. Hierarchical relations, which establish relations of hypernymy/hyponymy and meronymy, and 3. Relationships by association, which specify any other contextual, semantic or use relationship (Reichhold et al. 2012; Müller-Riedlhuber 2017).

The alignment of a topic of knowledge against the DISCO II thesauri requires a similarity between the topic and a taxonomic level of thesaurus tree. Figure 1 presents three cases of similarity with DISCO II thesaurus, where topics belong to the same subtree within the thesaurus and, therefore, have the same upper hierarchical level in the tree. Thus, for example, topics such as "network computing" and "parallel computing", besides having a lexical similarity (by the word computation), have a relationship of meronymy because they share the same subtree within the thesaurus (case 1). This is also the case between "Geoinformatics" and "Geographic data processing", which have a synonymy relation (case 2), and for "database analysis" and "data modeling", there is a relation of hyperonymy/hyponymy because these topics are part of the subtree corresponding to "knowledge of databases" (case 3). Consequently, to achieve the alignment of two knowledge topics, the first step is to find that topic of the tree whose lexical similarity for each topic is high, and then determines the degree of similarity between the subtrees of each one of them.

To perform the alignment of skills topics, there is a thesaurus of synonyms built on the basis of Bloom's taxonomy (Anderson et al. 2001), proposed in (Ortiz Sánchez 2016), which contains 6 cognitive levels (knowledge, understanding, application, analysis, synthesis, evaluation), 255 verbs associated with each cognitive level, and approximately 800 synonyms related to each verb. The relationships between the verbs of this thesaurus correspond to the belonging to a cognitive level, either by its inclusion in the set of related verbs, or in the set of synonyms.

In the same way, the alignment of a topic of skill against the BLOOM thesauri requires obtaining a similarity of the topic with the taxonomic levels of the thesaurus. Figure 2 presents two cases of similarity of skill topics in the alignments with the BLOOM thesaurus. As we can see, there is a synonymic relation between "articulate" and "compose", since regardless of whether they belong to different groups of related

[1] DISCO II, available online in http://disco-tools.eu/disco2_portal/projectInformation.php

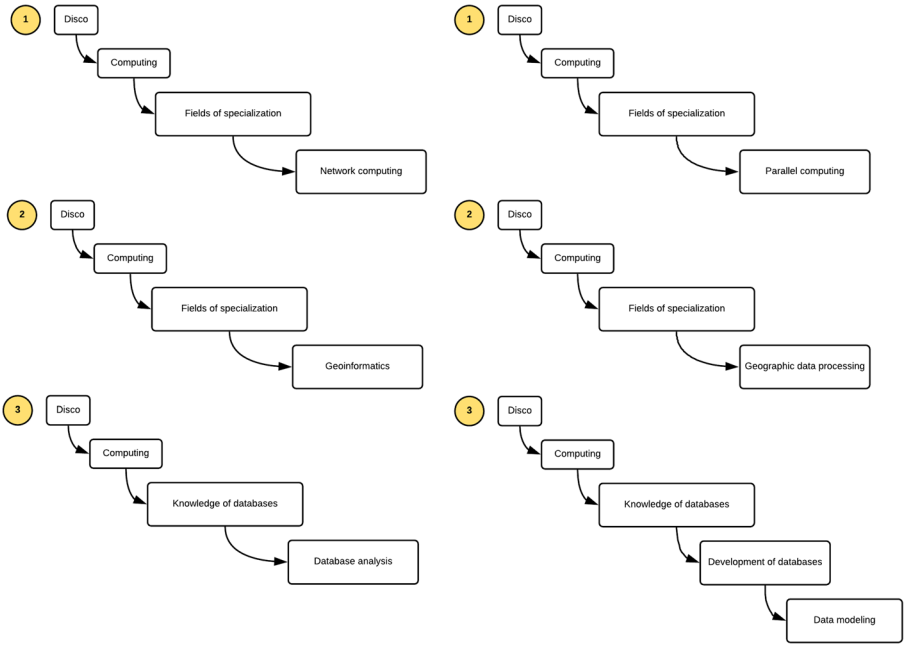

Fig. 1 Different types of topic similarity according to the DISCO II thesaurus, in cases of meronymy (1), synonymy (2) and hyperonymy/hyponymy (3)

verbs (assemble and write), they are under the cognitive level "Synthesis", which determines that both skill topics are similar (case 1). In the same way, there is a relation of similarity between "program" and "develop", because they are under the cognitive level "Application", although they do not belong to the same set of related verbs or synonyms (case 2). In summary, the alignment of two skill topics is obtained, firstly by finding that group of the thesaurus in which each topic is found, and then determining if they have the same cognitive level, or that skill topic, because they are at a higher cognitive level, covers another.

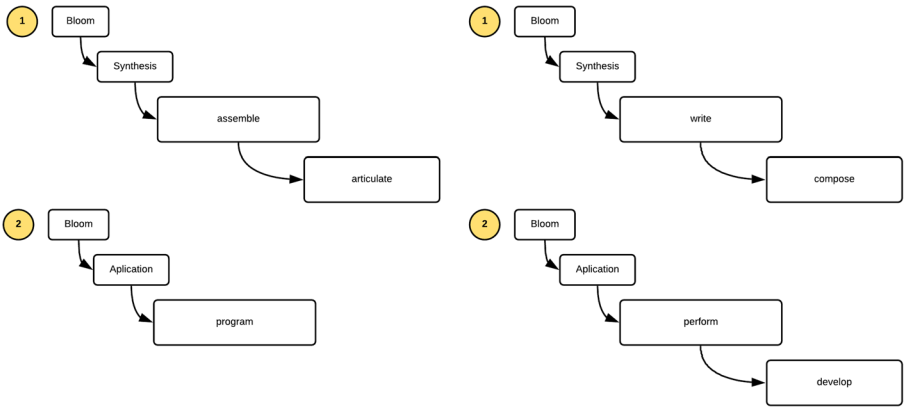

Fig. 2 Cases of synonymy for different skill topics according to the BLOOM thesaurus

Architecture

Figure 3 shows the general architecture for determining competencies in job advertisements, which consists of the following phases: the first phase performs the alignment of topics against thesauri, obtaining as a result measures of similarity (lexical and semantic); and, the second phase corresponds to the alignment of the profiles based on the similarity measures obtained in the first phase. There is an initial step, for the pre-processing of the texts from the Web (see Fig. 4), in order to obtain the knowledge and skill topics, which is based on a linguistic analysis that uses linguistic patterns formed by sequences of words with specific grammatical categories, defined in (González-Eras and Aguilar 2015).

The pre-processing is based on the approach defined in (González-Eras and Aguilar 2015), which uses competence concepts and about its elements (skills and knowledge), applied in each domain (academic and professional), to describe them. Then, it defines logical descriptions to characterize their patterns. These patterns are used during the pre-processing step of our architecture (Fig. 4). Particularly, a pre-processing is carried out, where HTML tags (headers, numbers, dates, metadata) are deleted, leaving only those texts that are inside type tags <p > </ *p*>. Then, a morpho-lexical analysis is carried out to extract sentences and words (tokens), which are normalized (capitalization, lemmatization, etc.) (Manning et al. 2009) and labeled with a grammatical category (Faria et al. 2014). Finally, in the analysis of patterns, the patterns are applied to the text in order to recognize the topics of skill and knowledge (González-Eras and Aguilar 2015). The whole process is supported by NLP tools that offer libraries for the development of each step in different languages (González-Eras and Aguilar 2019). Some examples of this process are found in section A of the "Experimentation" section (see Figs. 5 and 6).

With respect to the alignment of topics against thesauri, it is carried out based on the lexical and semantic similarities defined below, which determine the similarity between the skills and knowledge extracted from the academic profiles and job advertisements

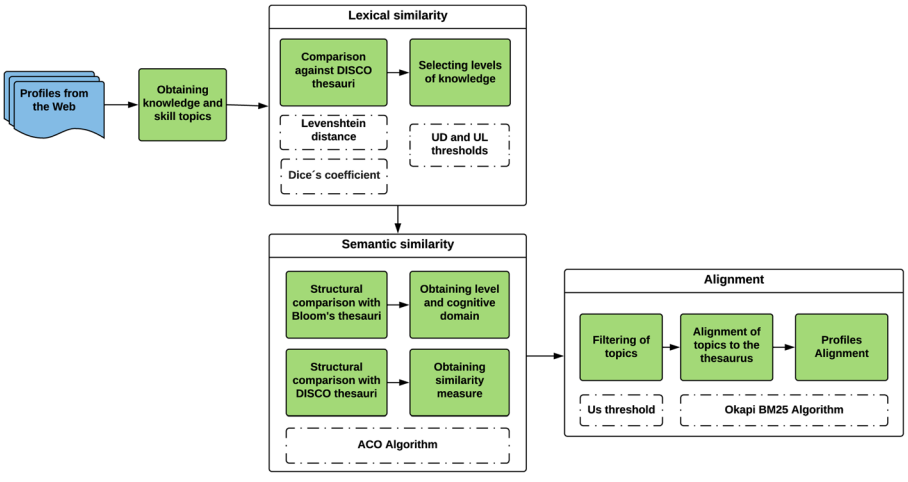

Fig. 3 Profile alignment architecture

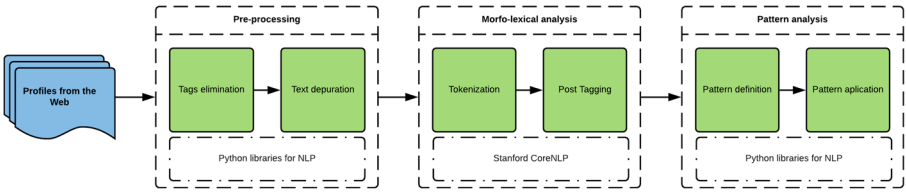

Fig. 4 Phase of obtaining of knowledge and skill topics

with these thesauri (see Fig. 3). Finally, the aligned topics of the academic profiles and job advertisements with the thesauri, are now aligned between them. These phases are based on a set of definitions and algorithms, which are below defined in this section.

Statement 1 Let $X = \{(id_1,p_1,c),\cdots, (id_n,p_n,c)\}$ be a valid collection of professional profiles (job advertisements), where id_i is a profile identifier, p_i is a set of multidimensional features and c has a value as defined in Eq. (1).

$$c = \begin{cases} 1 \text{ if } p_i \text{ is an academic profile} \\ 2 \text{ if } p_i \text{ is a job profile} \end{cases} \quad (1)$$

Definition 1 *A p_i profile is a collection of phrases described according to Eq.(2),*

$$p_i = \{F_i \ldots F_n\} \quad (2)$$

Definition 2 *A phrase F_i is a collection of topics described according to Eq. (3),*

$$F_i = \{(H_i, C_i) \ldots (H_n, C_n)\} \quad (3)$$

Where a profile p_i can have one or more phrases and each phrase F_i can contain one or more knowledge topics C_i or skill topics H_i. Table 2 presents an example of this structure.

Likewise, for the DISCO and BLOOM thesauri the following definitions are made, which correspond to the structures presented in Figs. 1 and 2:

F	*c*	*H*			*C*		
Diseñador y administrador de sistemas de comunicación de datos (Designer and administrator of data communication systems)	1	Diseñador (Designer)	y (and)	Administrador (Administrator)	comunicación (communication)	de (of)	datos (data)
		NC	CC	NC	NC	SP	NC
Desarrollo de aplicaciones computacionales (Development of computer applications)	1	Desarrollo (Development)		de (of)	aplicaciones (applications)		computacionales (computer)
		NC		SP	NC		AQ
Interacción con bases de datos (Interaction with databases)	2	Interacción (Interaction)		con (with)	bases (bases)	de (of)	datos (data)
		NC		CC	NC	SP	NC
Conocimiento avanzado de Java (Advanced knowledge of Java)	2	Conocimiento (knowledge)		avanzado (advanced)	de (of)		Java
		NC		AQ	SP		NC

Fig. 5 Example of the data processing of the experiment

c	id	F	H	C
1	1	Diseñador y administrador de sistemas de comunicación de datos (Designer and administrator of data communication systems)	Diseñar (design) Administrar (manage)	Sistemas de comunicación de datos (data communication systems)
1	2	Desarrollo de aplicaciones computacionales (Development of computer applications)	Desarrollar (Develop)	aplicaciones computacionales (computer applications)
2	1	Interacción con bases de datos (Interaction with databases)	Interactuar (Interact)	bases de datos (databases)
2	1	Conocimiento avanzado de Java (Advanced knowledge of Java)	Conocer (know)	Java

Fig. 6 Excerpt from the experimental dataset

Statement 2 Let $D = \{(C'_1,n_1),\cdots, (C'_n,n_n)\}$ be a set of knowledge topics organized hierarchically, where a topic C' belongs to a level n.

Definition 3 *A topic C' may have associated a set of phrases F' described according to Eq. (4),*

$$C'_i = \{F'_i \ldots F'_n\} \tag{4}$$

Definition 4 *A phrase F'_i is a collection of topics described according to Eq. (5),*

$$F'_i = \{(H'_i, C'_i) \ldots (H'_n, C'_n)\} \tag{5}$$

Where a topic C'_i may have zero or more competence phrases F'_i, and each one of them may contain one or more knowledge topics C'_i or skill topics H'_i.

Table 2 Example of statement 1 structure

id_i	C	P
1	1	Abordar proyectos de automatización computacional. (Addressing computational automation projects.) Diseñador y administrador de sistemas de comunicación de datos control de hardware y software. (Designer and administrator of data communication systems, hardware and software control.) Abordar proyectos de automatización computacional, mandos de máquinas eléctricas. (Address computer automation projects, electrical machine controls.) Integrar equipos, en operación y mantenimiento de sistemas electrónicos. (Integrate equipment, in operation and maintenance of electronic systems.)
2	1	Desarrollo de aplicaciones computacionales. (Development of computational applications.) Diseño y manejo de base de datos estadísticos. (Design and management of statistical database.)
3	2	Interacción con bases de datos y lenguaje SQL. (Interaction with databases and SQL language.) Programar y desarrollar en lenguaje Java POO. (Program and develop in Java OOP language.) Conocimiento avanzado de Java. (Advanced knowledge of Java.) Conocimientos de JRE 5, 6 y 7 y de programación concurrente en Java. (Knowledge of JRE 5, 6 and 7 and concurrent programming in Java.)
4	2	Conocimiento en Lenguaje SQL. (Knowledge in SQL Language.) Análisis y diseño de base de datos. (Analysis and design of database.)

Statement 3 Let $B = \{(TH_1,n_1),\cdots,(TH_n,n_n)\}$ be a set of hierarchically organized skills topics, where a *TH* topic belongs to a level *n*.

On the other hand, for the implementation of the architecture, it is necessary to make the following definitions:

Definition 5 *A topic TH has associated a set of related verbs V, described according to Eq. (6),*

$$TH_i = \{V_i \ldots V_n\} \tag{6}$$

Definition 6 *A verb V_i has a collection of S synonyms described according to Eq. (7),*

$$V'_i = \{S_i \ldots S_n\} \tag{7}$$

A process of similarity analysis, which is defined by the following stages, uses these definitions:

Lexical Similarity

Lexical similarity calculation of knowledge topics is explained in Table 3. With this algorithm, a lexical similarity is established between each C_i knowledge topic (profiles) and each topic belonging to the DISCO C'_i thesaurus. For this, two measures are considered, the first one called $Dis_{lex}(C,C')$, which uses the

Table 3 Pseudocode of the calculation of the lexical similarity

Start
Input
Collection of academic and employment profiles according to the Statement 1
DISCO thesauri of knowledge topics according to Statement 2
Procedure
1. Creation of profiles dataset P according to the Statement 1 and Definitions 1 and 2
2. Creation of the DISCO thesauri tree of knowledge topics D according to Statement 2 and Definitions 3 and 4
3. For all profile dataset P
3.1. Get the profile knowledge topic (C_i)
3.2. For all the Tree of knowledge topics D
3.2.1 Get the knowledge topic on the D Tree (C'_j)
3.2.2. Calculate the editing distance between C_i and C'_j $Dis_{lex}(C_i,C'_j)$ according to Definition 7
3.2.3. If $Dis_{lex}(C_i,C'_j)$ is greater than the distance threshold U_D according to the Statement 4
3.2.3.1. Calculate the lexical similarity between knowledge topics C_i and C'_j $Sim_{lex}(C_i,C'_j)$ according to the Definition 8
3.2.3.2. If $Sim_{lex}(C_i,C'_j)$ is less than the lexical threshold U_L according to the Statement 5
3.2.3.2.1. Discard C_i
3.2.3.3 If not
3.2.3.3.1 Add topic C_i to the profile dataset P' according to the Statement 6
3.2.4If not
3.2.4.1Discard C_i

Levenshtein measure to determine the edit distance between the topics considered (Levenshtein 1966); and the second one called $Sim_{lex}(C,C')$, which uses the Dice's coefficient to determine the similarity between topics, according to the similarity of its character pairs (Alqadah and Bhatnagar 2011). These measures are described in the following definitions.

Definition 7 *The editing distance between two topics C* and *C' is given by the number of character changes that must be made so that the topic C becomes the topic C' (see Eq. (8)) (Levenshtein 1996)*,

$$Dis_{lex}\left(C,C^{'}\right)=\begin{cases}\max\left(C,C^{'}\right)\ If\ Dis_{lex}\left(C,C^{'}\right)=0\\ \min\left(C,C^{'}\right)\ If\ Dis_{lex}\left(C,C^{'}\right)>0\end{cases} \tag{8}$$

Then, the value of the measure is maximum when the number of changes is zero (*C* and *C'* are equal), and is minimum otherwise.

Definition 8 *The lexical similarity between two topics C and C' is twice the number of pairs of characters that are common to both topics, divided by the sum of the number of pairs of characters in the two topics (see* Eq. 9*)*,

$$Sim_{lex}\left(C,C^{'}\right)=\frac{2\ x\ |pairs(C)\cap pairs\left(C^{'}\right)|}{|pairs(C)|+|pairs\left(C^{'}\right)|} \tag{9}$$

Then, for each pair of topics is compared their characters, and a similarity value between zero and one is obtained, where zero represents no similarity and one represents high similarity (Alqadah and Bhatnagar 2011).

Statement 4 A distance threshold U_D equal to four is established, which corresponds to the minimum edit distance that can exist between *C* and *C'* to consider that they have a similarity. The U_D value was defined by observing the result of the similarity calculation in 100 cases, based on the work done in (Dijkman et al. 2011). Table 4 presents an

Table 4 Units for magnetic properties

C	C'	Dis_{lex}(C,C')	Sim_{lex}(C,C')
Sistemas (systems)	Sistemas operativos (operating systems)	11	0.76
Sistemas (systems)	Sistemas distribuidos (distributed systems)	13	0.76
Base de datos (database)	Base de datos estadísticas (Statistical database)	13	0.82
Desarrollo de base de datos (Development of database)	Desarrollar base de datos (Develop database)	4	0.9

example of the calculation of the lexical similarity of the topics of the dataset, based on the two measures mentioned. As shown, the use of the two lexical measures increases the coverage of similar topics within the dataset. For example, the topics "operating systems" and "distributed systems" would have a low similarity in relation to the topic "systems" ($Dis_{lex} > 4$), if only the lexical distance would be taken into account as a comparative measure, the same happens with the topics "database" and "statistical database".

Statement 5 A U_L threshold equal to 0.4 is established, which corresponds to the minimum lexical similarity that may exist between C and C' to consider that they have a similarity. The U_L value was defined by observing the similarity calculation in 100 cases, based on the work done in (Dijkman et al. 2011; Van Dongen et al. 2013).

Statement 6 Let P' = $\{(c,id_1,C_1,H_1,n),\cdots, (c,id_n,C_n,H_n,n)\}$ be a valid topic dataset, where c indicates the type of profile according to (1). id_i is the identifier of the profile, C_i represents the topic of knowledge profiles, H_i is the ability related to the topic, and n is the tree level D where the maximum similarity value of the topic is found, which fulfills the thresholds defined in the statements 4 and 5. Table 11 presents an example of this structure, which is the result of the lexical similarity phase.

Semantic Similarity

Semantic similarity calculation of profile topics is explained in the macro algorithm of Table 5.

The procedure begins with a pair's alignment analysis of selected topics by their lexical similarity against the DISCO II Thesauri (Müller-Riedlhuber 2009), by means of a scheme proposed in (González-Eras and Aguilar 2015), in which, for each pair of topics C and C', the similarity of their ancestors, brothers and sons in the thesaurus is verified (Mendonza et al. 2015).

Structural Comparison with DISCO Thesaurus

Definition 9 *The semantic similarity of two topics* C and C' *is given by the sum of the similarities of ancestors, siblings and children of topic* C, *divided by 3. Then, for each pair of topics a similarity value is obtained in the range of zero to one, considering that zero represents no similarity and one represents high similarity (see Eq. (10)).*

$$Sim_{sem}\left(C, C^{'}\right) = \frac{SA\left(C, C^{'}\right) + SD\left(C, C^{'}\right) + SS\left(C, C^{'}\right)}{3} \tag{10}$$

Here it is presented each one of the measures:

Table 5 Pseudocode of the calculation of the semantic similarity

Start

Input

Dataset of knowledge topics P' according to Statement 6

DISCO thesauri tree of knowledge topics according to Definitions 3 and 4

BLOOM thesauri of skill topics according to Statement 3

Procedure

1. Creation BLOOM thesauri tree of skill topics B according to Definitions 5 and 6
2. For all profile Dataset P'
   - 2.1. Get the knowledge topic (C_i)
   - 2.2. For all the Tree of knowledge topics D
     - 2.2.1. Calculate the semantic similarity between C_i and C'_j Sim_{sem} (C_i, C'_j) according to Definition 9
     - 2.2.2. Get the root topic of the thesauri DISCO subtree m_d of maximum Sim_{sem} (C_i, C'_j) according to Statement 7
     - 2.2.3. Add to knowledge topic dataset TC C_i, max Sim_{sem} (C_i, C'_j), m_d, id and c according to Statement 9
   - 2.3. End For
3. End For
4. For all P' Dataset
   - 4.1. Get the skill topic (H_i)
   - 4.2. For all B Tree
     - 4.2.2. Get the root topic of the thesauri BLOOM subtree m_b where the skill topics (H_i, H'_j) are found according to Statement 8
     - 4.2.3. Add to the skill topic dataset TH H_i, m_b, id and c according to Statement 10
   - 4.3. End For
5. End For

Definition 10 *The similarity of topics C and C' will be proportional to the similarity of their Ancestors concepts. In this case, the average of the maximum similarities of each ancestor of topics C* and *C' is considered (see Eq. (11)),*

$$SA\left(C, C'\right) = \frac{1}{n}\sum_{i=1}^{n} \max\left(Sim\left(Anc_i(C), Anc_1\left(C'\right)\right), \ldots, Sim\left(Anc_i(C), Anc_n\left(C'\right)\right)\right) \tag{11}$$

Where:

$Anc_i\ (C)$	ancestor *i* of topic *C*.
$Sim(Anc_i(C),Anc_j(C'))$	measure of similarity between ancestors of topics *C* and *C'*, according to Definition 8.
n	maximum level of lexical similarity between *C* and *C'*.

Definition 11 *The similarity of topics C* and *C' will be proportional to the similarity of the siblings. In this case, the average of the maximum similarities of the siblings of topics C and C' is considered (see Eq. (12)),*

$$SS\left(C, C'\right) = \frac{1}{n}\sum_{i=1}^{n} \max\left(Sim\left(Sin_i, Sin'_1\right), \ldots, Sim\left(Sin_i, Sin'_n\right)\right) \tag{12}$$

Where:

Sin_i	corresponds to the i brother of topic C.
Sin'_j	corresponds to the j brother of topic C'.
$Sim(Sin_i, Sin'_n)$	the measure of similarity between the siblings of topics C and C' according to Definition 8.
n	maximum level of lexical similarity between C and C'.

Definition 12 *The similarity between two topics C and C' will be proportional to the similarity of their direct descendants. In this case, the average of the maximum similarities of the children of topic C with the children of topic C' is considered (Eq. 13),*

$$SD\left(C, C'\right) = \frac{1}{n} \sum_{i=1}^{n} \max(Sim(Des_i, Des'_1), \ldots, Sim(Des_i, Des'_n)) \quad (13)$$

Where:

Des_i	corresponds to the son of topic C.
Des'_j	corresponds to the son of topic C'.
$Sim\ (Des_i, Des'_j)$	the measure similarity between the children of topics C and C' according to Definition 8.
n	maximum level of lexical similarity between C and C'.

Statement 7 It is considered that m_d represents the root topic of the subtree of the DISCO thesauri, where topics C and C' reach a maximum similarity value.

Structural Comparison with the Bloom Thesaurus

The process of the skill topics structural comparison includes its comparison with the skill topics of the BLOOM thesaurus of Statement 3, which implies identifying the cognitive level that is the root of the subtree where the compared topics are located, for which purpose it is posed the following statement:

Statement 8 It is considered that m_b represents the cognitive level that is the root of the subtree of the BLOOM thesaurus, where skill topics H and H' are found. Table 6 presents examples of semantic similarity on skill topics of the profiles. As we can see, there is a similarity relation between "innovate and advise" (lines 3 and 4), since both topics belong to the same subtree of the BLOOM thesaurus, whose root corresponds to cognitive level "Apply", so that a relation of synonymy is established between them. For the other topics, there is a relation of non-similarity, because they do not share the same subtree in the thesaurus.

Table 6 Comparison of skill topics with the BLOOM Thesaurus

Skill topic	m_b	Equivalence
H: definer (define)	Analizar (analyze)	No (No)
H': plantear (propose)	Aplicar (apply)	No (No)
H: innovar (innovate)	Aplicar (apply)	Si (Yes)
H': asesorar (advise)	Aplicar (apply)	Si (Yes)
H: definer (define)	Analizar (analyze)	No (No)
H': gestionar (manage)	Crear (create)	No (No)

For results registration of topic semantic similarity, the following statements are made:

Statement 9 Let TC = {(c, id_1, C_1, Ms_1, md_1), ···, (c, id_n, C_n, Ms_n, md_n)} be a valid knowledge topic dataset, where *c* indicates the profile type according to (1). id_i is the profile identifier, C_i represents the profile knowledge topic; *Ms* is the maximum measure of semantic similarity $Sim_{sem}(C_i,C'_i)$, obtained according to Eq. (10), and *md* is the root topic of the DISCO thesaurus subtree according to Statement 7. Table 12 presents an example of this structure, which is the result of the semantic similarity phase of knowledge topics. Statement 10. Let TH = {(c, id_1, H_1, Ms_1, mb_1), ···, (c, id_n, Cn, Ms_n, mb_n)} be a valid skill topic dataset, where c indicates the profile type according to (1). id_i is the profile identifier, H_i represents the profile skill topic; Ms is the maximum measure of semantic similarity $Sim_{sem}(H_i,H'_i)$, obtained according to eq. (10), and mb is the root topic of the BLOOM thesaurus subtree according to Statement 8. Table 13 presents an example of this structure, which is the result of the semantic similarity phase of skill topics.

Alignment

The profile alignment process is performed according to the root topic of the thesaurus subtree where was found the similarity relation of knowledge and skill topics in the semantic similarity phase. For this purpose, Table 7 explains the process in the following macro algorithm:

The process of alignment of the profiles begins with the filtering of knowledge and skill topics, using the *Us* threshold, which allows selecting those topics with a greater measure of semantic similarity according to Eq. (10) Then, for each one of the remaining knowledge and skill topics, the frequency of them in the profiles is calculated. With this frequency, the position of the profiles is established according to the root topic of the thesaurus tree to which they belong (Jones et al. 2000). These measures are described in the following definitions.

Statement 11 A threshold equal to 0.45 is established, which corresponds to the minimum measure that can exist between *C* and *C′* to consider that they have a semantic similarity. U_s is equal to 0.45 The value of U_s was defined by observing the result of the similarity calculation in 100 cases, based on the work done in (González-Eras and Aguilar 2015). Table 8 presents examples of the calculation of the semantic

Table 7 Alignment phase pseudocode

Start
Input
Knowledge topic dataset TC according to Statement 9
Dataset of TH skill topics according to Statement 10
Procedure 1. For the whole TC knowledge topic dataset 1.1. Get the knowledge topic (C_i) 1.2. If the measure of semantic similarity of the topic C_i Ms (Eq. (11)) is smaller than the threshold of semantic similarity Us according to Statement 11 1.2.1. Discard the knowledge topic C_i 1.2.2. Discard the topic of skill H_i associated with the topic C_i 1.3. End IF 2. End For 3. For the whole TC knowledge topic dataset 3.1. For each profile id_i 3.1.1. Calculate the relevance of the id_i profile Score (id_i, md_i) according to Definition 14 3.2. End For 4. End For 5. For the whole TH skill topic dataset 5.1. For each id_i profile 5.1.1. Calculate the relevance of the id_i profile Score (id_i, mb_i) according to Definition 14 5.2. End For
End

similarity on the knowledge topics of the profiles. As is shown, the calculation process allows obtaining those pairs of topics that are semantically similar, according to the structure of the DISCO thesaurus tree, associating this similarity with the root topic of the subtree that they share in the thesaurus. The threshold of similarity ($U_s > 0.45$) allows detecting the topics of knowledge that present a greater semantic similarity.

Definition 14 *The relevance value of a Profile consists of the position of the profile within the collection of profiles analyzed, based on the knowledge and skill topics it*

Table 8 Instance Comparison using the DISCO Thesaurus

C	C′	$Sim_{sem}(C,C')$	md
Sistemas de comunicación (Comunication system)	sistemas de información y comunicación (Information and communication systems)	0.67	instalación y configuración TI (IT installation and configuration)
Control de hardware (Hardware control)	configuración de hardware (hardware configuration)	0.69	instalación y configuración TI (IT installation and configuration)
Análisis de sistemas (System analysis)	Análisis de TI (IT Analysis)	0.61	Análisis TI (IT Analysis)
Sistemas informáticos (Computer systems)	Depuración de sistemas (System debugging)	0.8	Administración TI (IT Administration)

contains. In particular, the relevance value of an Id_i profile according to the topic of the md_i thesaurus (knowledge or skill), is given by Eq. (14). This is a measure used in information retrieval, known as Okapi BM25, which orders by relevance the documents in function of the topic they contain (Robertson and Zaragoza 2009). The classical metric TF-IDF takes into account the frequency of occurrence of a topic within the collection of documents (Jones et al. 2000), but Okapi BM25 is more sensitive because also takes into account the length of the documents (Yuanhua and Zhailk 2011).

$$Score(id_i, md_i) = \sum_{i=1}^{n} IDF(md_i).\ \frac{f(md_i, id_i).(k_1+1)}{f(md_i, id_i) + k_1.\left(1-b+b.\frac{|D|}{avgdl}\right)} \quad (14)$$

Where:

f(md$_i$, idi)	is the frequency of topical md_i in the id_i profile according to the definition 15.
| D |	is the number of topics (length of profile id_i).
avgdl	is the average length of the profiles that make up the collection.
K_1 and b	adjustable parameters of the function *Score(id$_i$,md$_i$)*[2] to the set of profiles of the specific characteristics (frequency of topics and length of the document, respectively) (Robertson and Zaragoza 2009).
IDF (md$_i$)	is the weight given to topical md_i in the collection of profiles, according to the definition 16.
n	number of profiles in the collection.

Definition 15 *The frequency of appearance of a topic consists of the number of knowledge topics roots of the sub-trees of the thesaurus that contains the profile. The frequency of occurrence of the topic md_i* in *id_i is given by Eq. (15),*

$$f(md_i, id_i) = \sum_{i=1}^{n} md_i \quad (15)$$

Where *n* represents the total of md topics found in the id_i profile (Manning et al. 2009).

Definition 16 *The weight of a topic md_i IDF(md_i) is given by the inverse frequency of the same in relation to the collection of profiles, which is presented in the following Eq. (16),*

$$IDF(md_i) = log\frac{N-md_i+\delta}{md_i+\delta} \quad (16)$$

Where *N* is the number of profiles in the collection, *n(md_i)* is the number of profiles that contain the topic md_i, and δ is a parameter of adjustment to the weight given to a topic, according to the characteristics of its frequency in the collection of profiles and the length of the documents (Yuanhua and Zhailk 2011).

[2] Text Retrieval Conference, disponible en http://trec.nist.gov/

Experimentation

Processing of Experimental Data

The general architecture for determining competencies of the "Architecture" section has been automated. In this section, we study its behavior. For the experiment, 35 documents in Spanish were taken as input: 20 academic profiles, obtained from university portals (id_1, ..., id_{20}), and 15 job advertisements, obtained from internet employment portals (id_{21}, .. id_{35}). From each profile, text extracts were selected that were under sections, such as description, objectives, competencies, skills, knowledge. In these sentences, there are elements of competence, such as skills and knowledge, of which we can see examples in Table 1.

The first step is the pre-processing of the texts to obtain the knowledge and skill topics, according to the procedure indicated in the general architecture (see "Characterization of the Competence Concept" section). It starts with the development of a linguistic analysis, to recognize the instances of knowledge and skill, based on linguistic patterns (González-Eras and Aguilar 2015). Figure 5 presents an example of the analysis for the first sentence of the profiles represented in Table 2, where the knowledge instances are recognized according to patterns, which are formed by the noun, preposition or adjective sequences ([NC], [NC-SP-NC], [NC-NC], [NC-AQ])[4]; while skill instances by patterns with verb, noun, preposition or conjunction sequences ([VMN], [NC-SP], [NC-CC-NC])[3] (González-Eras 2017).

As a result, 93 instances of knowledge and 70 instances of skill were detected in the academic profiles, while in the job advertisements 204 instances of knowledge and 96 of skills were detected. In Fig. 6, an extract of the structure of the dataset is presented, which was organized, as indicated in statement 1 and definitions 1 and 2.

Semantic Sources

For the present experiment, Table 9 presents the root topics of the sub-trees of the DISCO thesaurus against which the profiles are aligned, which correspond to sub-areas of Computer Science and Computer Science. In the same way, Table 10 presents the root topics of the BLOOM thesaurus sub-trees, which correspond to the cognitive levels defined in Bloom's taxonomy.

Profiles Alignment

Phase 1: Lexical Similarity In this phase, the lexical similarity between the knowledge topics of the profiles and the DISCO Thesaurus is sought through the similarity measures presented in Eqs. (8) and (9) of the macro algorithm of Table 3.

It is observed that the calculated distance between instance C and C', according to Definitions 7 and 8 and the U_L threshold (statement 5), allows identifying the topic in the DISCO thesaurus, with which the topics of the profiles have a greater lexical similarity.

[3] CONLL Format, VMN: verb, CC: conjunction.

Table 9 Definition of the roots of the subtrees of the thesaurus Disco II

	md
Tc1	Instalación y configuración TI (IT installation and configuration)
Tc2	Desarrollo de software (Software development)
Tc3	Campos de especialización en TI (Special IT fields)
Tc4	Consultoría de TI (IT consulting)
Tc5	AnálisisTI (IT analysis)
Tc6	Programación (Programming)
Tc7	Conocimientos de bases de datos (Database knowledge)
Tc8	Sistemas operativos (Operating systems)
Tc9	Gestión de proyectos TI (IT project management)
Tc10	Administración de TI (IT administration)
Tc11	Internet y multimedia (Internet and multimedia)
Tc12	Informática (Computing)
Tc13	Seguridad de la información (Information security)
Tc14	Soporte TI (IT support)
Tc15	Tecnología de red (Network technology)

This is important for the development of the next phase, because we know the level of the thesaurus where this topic is located, and thus, we obtain the subtree on which the calculation of the taxonomic measure will be made. Table 11 shows an extract of the result of the calculation of the lexical similarity for the instances of the dataset of Fig. 5, according to Statement 6. As is seen in the table, the use of the measurements with the threshold allows identifying that subtree that presents a greater possibility of relating to the context of the topic. For example, "statistical databases - databases" and "Java – Java Language" share the same context, so there is a semantic similarity between these topics.

Phase 2: Semantic Similarity In this phase, we look for the semantic similarity between knowledge topics and the DISCO thesaurus according to Eq. (10); and, the similarity between the topics of skill and the thesaurus BLOOM according to the Statement 8. For that, the macro algorithm of Table 5 is invoked.

It is observed that the measure calculated between the instances C and C', according to definition 9, allows identifying the general topic of the DISCO thesaurus, with which the topics of the profiles have a greater similarity. Table 12 shows an extract of the

Table 10 Definition of the topics of the subtrees of the thesaurus Bloom

	mb
Th1	Conocimiento (Knowledge)
Th2	Comprensión (Understanding)
Th3	Aplicación (Application)
Th4	Síntesis (Synthesis)
Th5	Creación (Creation)
Th6	Evaluación (Evaluation)

Table 11 Calculation of the lexical similarity for topics of the profiles

c	id	C	C′	N	$Sim_{lex}(C,C')$
1	1	Sistemas de comunicación de datos (Data communication systems)	Comunicación de datos (Data communication)	3	0.6
1	2	Aplicaciones computacionales (Computer applications)	Aplicaciones informáticas (Computer applications)	3	0.7
2	1	Bases de datos estadísticas (Statistical data bases)	Base de datos (Databases)	2	0.9
2	1	Java	Lenguaje Java (Java language)	2	0.8

result of the calculation of the semantic similarity for the topics of the dataset of Fig. 5, according to Statement 9. For example, computer applications, computer projects, SQL language, Java language and Java have a relation of similarity with the general topic of the thesaurus "programming", which confirms that they belong to the same context, in this case Programming. In the same way, it is good to note here that, although the values of M_s are close to the threshold $U_s > 0.45$, there is a fairly clear similarity between these topics. In addition, the similarity value of the Java topic shows that the topic exists as it is written in the DISCO thesaurus; this is the reason why the Java language topic obtains a similarity value of 0.58 (lower).

In the same way, Table 13 presents the calculation of the similarity measure for instances H and H', using Statement 8. As is shown, according to the calculated measure, the administrator, direct, interaction and planning instances have a semantic relationship with the general topic "Synthesis", according to the BLOOM thesaurus. In the case of the topics of ability to develop and program, it is evident that there is a relation of synonymy within the context of "Application".

Table 12 Calculation of the semantic similarity for the knowledge topics of disco thesaurus

c	Id	C	Ms	md
1	1	Sistemas de comunicación de datos (Data communication systems)	0.67	Instalación y configuración TI (IT installation and configuration)
1	1	Actividades de especificación (Specification activities)	0.76	Desarrollo de software (Software development)
1	2	Aplicaciones computacionales (Computer applications)	0.59	Programación (Programming)
1	3	Proyectos informáticos (IT projects)	0.53	Programación (Programming)
2	1	Bases de datos (Databases)	0.7	Conocimientos de base de datos (Database knowledge)
2	1	Lenguaje SQL (SQL language)	0.61	Programación (Programming)
2	1	Lenguaje Java (Java language)	0.58	Programación (Programming)
2	1	Java	1	Programación (Programming)

Table 13 Calculation of semantic similarity for topics of skill with Bloom

c	Id	H	Ms	mb
1	1	Administrador (Administrator)	0.7	Síntesis (Synthesis)
1	1	Dirigir (Lead)	0.8	Síntesis (Synthesis)
1	2	Desarrollo (Developing)	0.8	Aplicación (Application)
1	3	Planificar (Plan)	1	Síntesis (Synthesis)
2	1	Interacción (Interaction)	0.7	Síntesis (Synthesis)
2	1	Interacción (Interaction)	0.7	Síntesis (Synthesis)
2	1	Programar (Program)	0.9	Aplicación (Application)
2	1	Conocer (Know)	1	Evaluación (Evaluation)

Phase 3: Alignment In this phase, using the previous similarity measures, we determine the alignment of profiles of the collection (their knowledge and skill topics) with the thesauri. With the aligned topics, we establish the topics around which the documents (academic profiles and job advertisements) relate between them, and those in which they have no relationship. For that, the algorithm of Table 7 is invoked. At the follows is given an example of this process on 3 documents: id_1, id_2 and id_{21}. The results of the alignment phase of the entire collection of documents are presented in "Discussion about the Obtain Result" section.

Table 14 and Fig. 7 show the result of the alignment of the profiles based on the knowledge topics, according to Definition 14, considering that k = 1.2 b = 0.75 and δ = 1. It is observed that the profiles id_2 and id_{21} are aligned around the topics Tc6 and Tc7 (programming and knowledge of databases respectively), being id_{21} the one that presents a greater value of relevance in relation to the topic "programming" (0.62 versus 0.59), while id_2 has a higher relevance value in the topic "knowledge of databases (0.41 versus 0.18). There is also an alignment between id_1 and id_2 around the topic Tc3 (fields of specialization in IT), where id_1 has the highest relevance value (0.29 against 0.16). The previous results indicate that the academic profile id_2 covers

Table 14 Calculation of the alignment of profiles and topics of knowledge

Topics	$n(md_i)$			$IDF(md_i)$			$Score(id_i,md_i)$		
	c		Total	c		Total			
	1	2		1	2		id_1	id_2	id_{21}
Tc1	3	1	4	0.802	1.176	0.923	1.33		
Tc2	6	5	11	0.465	0.415	0.436	0.54		
Tc3	12	4	16	0.082	0.528	0.235	0.29		0.16
Tc5	3	5	8	0.802	0.415	0.595			0.4
Tc6	3	9	12	0.802	0.087	0.391		0.59	0.62
Tc7	5	10	15	0.556	0.021	0.271		0.41	0.18

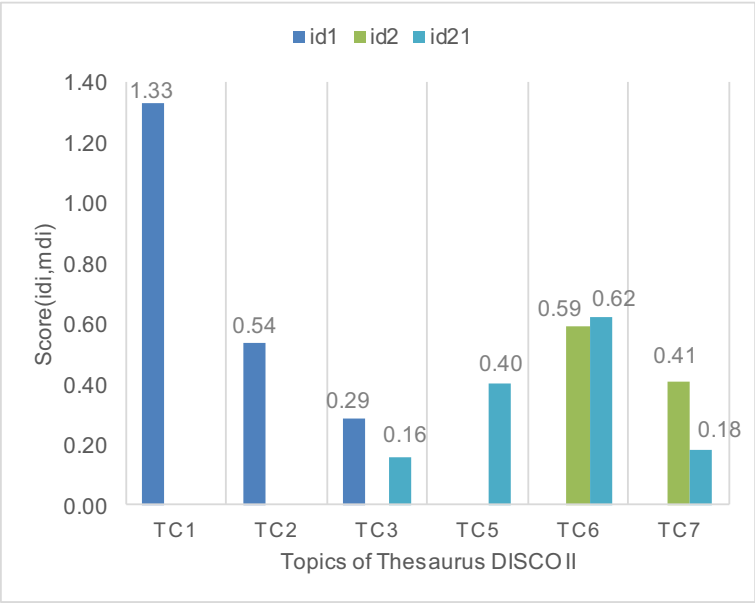

Fig. 7 Alignment of profiles id_1, id_2 and id_{21} according to knowledge topics

the requirements of the job advertisement id_{21}, it is not the case of id_1 that does not present any alignment with id_{21}. In addition, it is clear the high relevance value that reaches id_{21} in the topic Tc6 (programming), gives a first notion of feedback from the work context to the academic context, emphasizing the importance that companies give to this topic within their job offers.

On the other hand, the relevance value of id_1 in the topic Tc1 (installation and IT configuration) exceeds the value of 1 (1.33), because the number of profiles containing the topic Tc1 within the collection ($n(md_i)$) is low, with respect to the other topics (Tc1 is presented in 3 academic profiles and 1 job advertisement). Consequently, the relevance equation gives it a greater weight ($IDF(md_i)$ total of 0.923).

Table 15 Calculation of the alignment of profiles and topics of skill

Topics	$n(md_i)$			$IDF(md_i)$			$Score(id_i,md_i)$		
	c		Total	c		Total			
	1	2		1	2		id_1	id_2	id_{21}
Th3	12	6	18	0.08	0.28	0.30	0.46	0.41	0.12
Th4	1	4	4	1.32	0.57	0.86			0.38
Th5	18	13	31	0.07	0.06	0.06	0.10	0.11	0.10
Th6	8	13	21	0.31	0.06	0.23	0.23		0.30

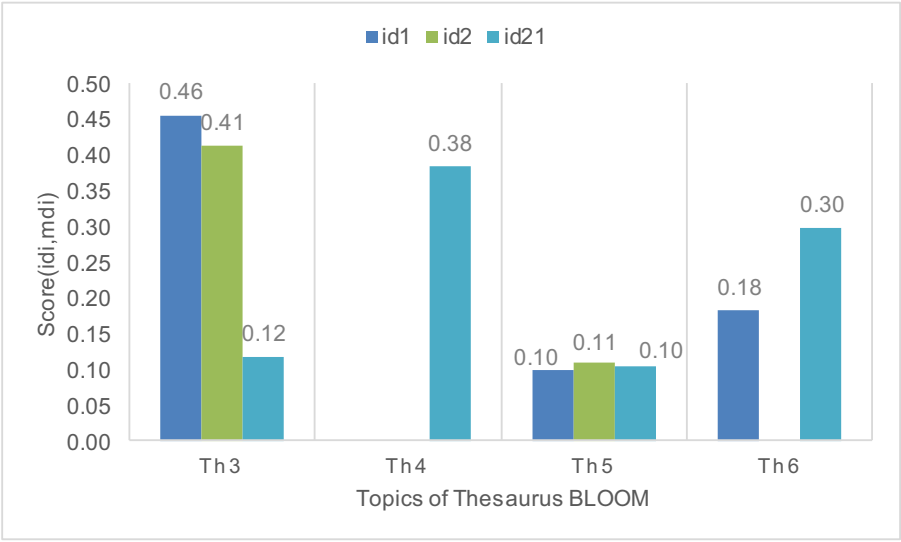

Fig. 8 Alignment of profiles id_1, id_2 and id_{21} according to skill topics

In the same way, Table 15 and Fig. 8 present the alignment of the profiles based on the topics of skill, according to Definition 14, considering $k_1 = 1.2$ $b = 0.75$ and $\delta = 1$.

It is observed that the profiles id_1, id_2 and id_{21} are aligned, around the topic Th3 (application), being id_1 the one that presents a greater value of relevance (0.46 against 0.41 and 0.12), which indicates that the academic profiles give great importance to the application of knowledge. There is also an alignment between id_1, id_2 and id_{21} around the topic Th5 (creation), where the three offers have very close relevance values (0.10, 0.11 and 0.10 respectively), which indicates that both two academic profiles cover to id_{21} in terms of the ability to create of knowledge. In the same way, id_{21} and id_1 are aligned around the topic Th6 (evaluation), highlighting this skill as a requirement of the job context, which is also present in the academic profile id_1 but in lower level (0.30 against 018, respectively). Finally, we identify that Th4 (synthesis) is a skill requested by companies, which is not considered in the academic profiles id_1 and id_2.

Discussion about the Obtained Results

Figures 9 and 10 present the results of the alignment of academic profiles and job advertisements (id_1…, id_{35}), with the knowledge topics of the DISCO II thesaurus (Tc1, …, Tc15). As is seen, the average of the documents of the collection focus on the topics: “software development” (Tc2), “IT specialization fields” (Tc3), “IT analysis” (Tc5), “programming” (Tc6), “knowledge of databases” (Tc7) and “operating systems” (Tc8). Some topics, like “IT project management” (Tc9), “IT administration” (Tc10), or “network technology” (Tc15), have a high alignment with one or several of the documents, but in general, their averages in the collection are low. Overall, the topics with the highest average of alignment in the documents of the collection are those

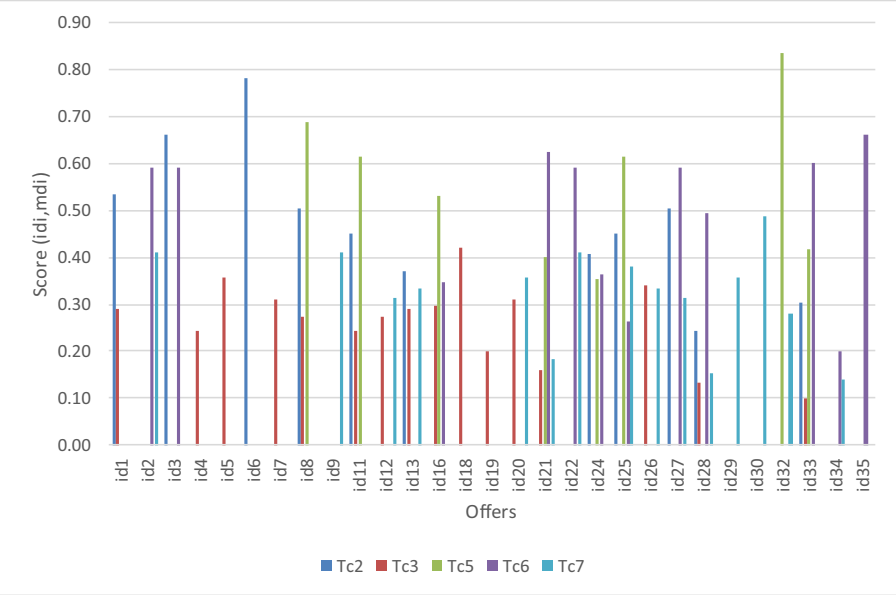

Fig. 9 Alignment of the documents with the knowledge topics (tc1 to tc7)) of the of the DISCO II thesaurus

comprised in the interval Tc1 to Tc8. With the other topics, the average is lower or does not exist alignment, as in the case of Tc4 (IT consulting).

Table 16 presents the alignment values of the profiles (documents) for the topics with the highest average of alignment. Summarizing, the collection of academic profiles and job advertisements present a tendency towards the first 8 topics of the

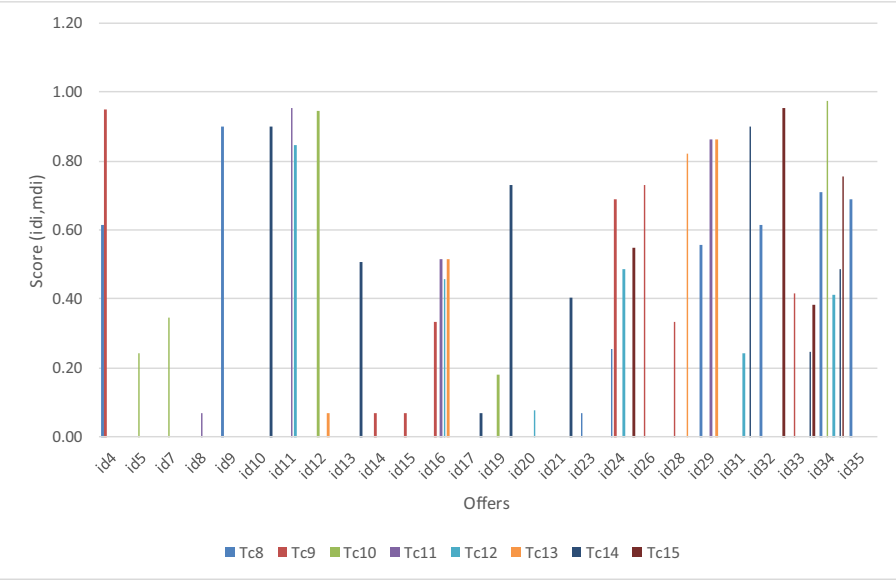

Fig. 10 Alignment of the documents with the knowledge topics (tc8 to tc15) of the DISCO II thesaurus

Table 16 Results of the alignment of profiles to the DISCO II thesaurus

	Tc1	Tc2	Tc3	Tc5	Tc6	Tc7	Tc8
id_1	0.33	0.54	0.29				
id_2					0.59	0.41	
id_3		0.66			0.59		
id_4			0.24				0.62
id_5			0.36				
id_6		0.78					
id_7			0.31				
id_8		0.50	0.27	0.69			
id_9						0.41	0.90
id_{11}		0.45	0.24	0.62			
id_{12}			0.27			0.31	
id_{13}	0.79	0.37	0.29			0.33	
id_{16}	0.29		0.30	0.53	0.35		
id_{18}			0.42				
id_{19}			0.20				
id_{20}			0.31			0.36	
id_{21}			0.16	0.40	0.62	0.18	
id_{22}					0.59	0.41	
id_{24}		0.41		0.35	0.36		0.25
id_{25}	0.62	0.45		0.62	0.26	0.38	
id_{26}			0.34			0.33	
id_{27}		0.50			0.59	0.31	
id_{28}		0.24	0.13		0.50	0.15	
id_{29}						0.36	0.56
id_{30}						0.49	
id_{32}				0.84		0.28	0.62
id_{33}		0.30	0.10	0.42	0.60		
id_{34}					0.20	0.14	0.71
id_{35}					0.66		0.69

DISCO II thesaurus. For example, for the topic Tc2, it can be seen that the documents have alignment values from higher to lower in the following way (see Table 16): id_6 (0.78), id_3 (0.66), id_1 (0.54), id_8 e id_{27} (0.5), id_{11} and id_{25} (0.45), id_{24} (0.41).

Figure 11 presents the results of the alignment of academic profiles and job advertisements, according to the skill topics of the BLOOM thesaurus. As is seen, the documents of the collection focus on the topics "application" (Th3), "synthesis" (Th4), "creation" (Th5) and "evaluation" (Th6), and the topics with the highest averages are Th3 and Th6. In the case of other topics, the average is smaller or does not exist alignment, as in the case of Th1 (knowledge) and Th2 (understanding). Table 17 presents the relevance value for the topics with the highest average of alignment in the documents.

It is said, then, that the collection of academic profiles and job advertisements present a tendency towards the topics Th3, Th4, Th5 and Th6 of the BLOOM

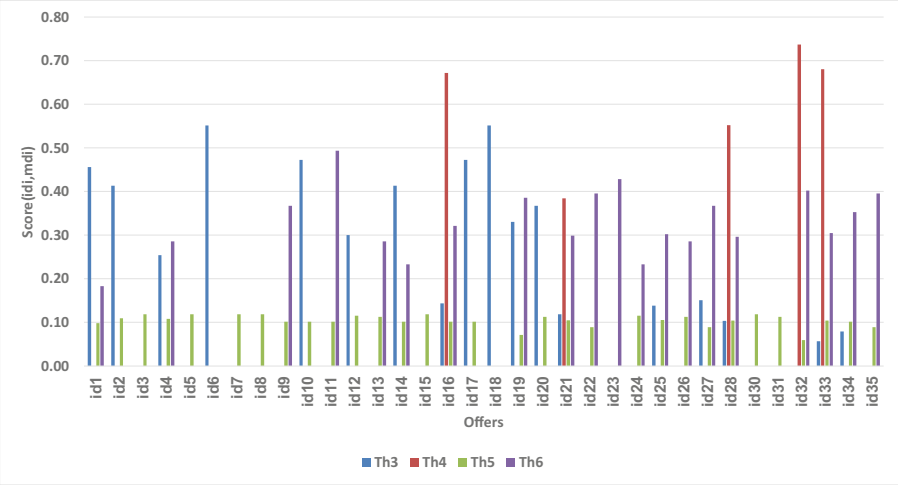

Fig. 11 Alignment of the profile collection according to skill topics

thesaurus, establishing a feedback between them according to their relevance value. For example, for the Th3 topic, it is seen that the profiles have values from higher to lower as (see Table 17): id_6 and id_{18} (0.55), id_{17} and id_{10} (0.47), id_1 (0.46), id_2 and id_{14} (0.41), id_{20} (0.37), id_{19} (0.33), id_{12} (0.30), id_4 (0.25), id_{27} (0.15), id_{16} and id_{25} (0.14), id_{21} (0.12), id_{28} (0.10), id_{34} (0.08) and id_{33} (0.06).

With the results obtained in this phase, we can establish that academic profiles and job advertisements are aligned to the same topic of the thesauri; and what is the strength of these alignments. In addition, feedback between them can be done; for example, what academic profiles cover the topics of knowledge required by job advertisements, or which universities have their academic profiles aligned to a work topic. In the same way, establishing what skills require job advertisements and what academic profiles can cover them. These results can be used in different contexts, like for the planning of professional careers, recruitment of personnel, among other domains.

Comparison with Other Works

The lexical similarity measures have been used in other works, such as editing distance (Levenshtein 1966), to find the similarity between concepts (Alqadah and Bhatnagar 2011), process names (Dijkman et al. 2011), or in the context of learning analytic approaches, to identify correlations between multimodal learning data, using different similarity metrics, such as the temporal similarity or the temporally relaxed similarity (Worsley and Blikstein 2018). In the same way, in the field of competencies, they have been used to establish hierarchical relationships between knowledge topics (Malzahn et al. 2013), and to classify documents using the Dice's coefficient as a measure of

Table 17 Results of the alignment of profiles to the BLOOM thesaurus

	Th3	Th4	Th5	Th6
id_1	0.46		0.10	0.18
id_2	0.41		0.11	
id_3			0.12	
id_4	0.25		0.11	0.29
id_5			0.12	
id_6	0.55			
id_7			0.12	
id_8			0.12	
id_9			0.10	0.37
id_{10}	0.47		0.10	
id_{11}			0.10	0.49
id_{12}	0.30		0.12	
id_{13}			0.11	0.29
id_{14}	0.41		0.10	0.23
id_{15}			0.12	
id_{16}	0.14	0.67	0.10	0.32
id_{17}	0.47		0.10	
id_{18}	0.55			
id_{19}	0.33		0.07	0.39
id_{20}	0.37		0.11	
id_{21}	0.12	0.38	0.10	0.30
id_{22}			0.09	0.40
id_{23}				0.43
id_{24}			0.12	0.23
id_{25}	0.14		0.11	0.30
id_{26}			0.11	0.29
id_{27}	0.15		0.09	0.37
id_{28}	0.10	0.55	0.10	0.30
id_{30}			0.12	
id_{31}			0.11	
id_{32}		0.74	0.06	0.40
id_{33}	0.06	0.68	0.10	0.30
id_{34}	0.08		0.10	0.35
id_{35}			0.09	0.40

similarity (Gomaa and Fahmy 2013); others to find the best contractors to perform the business process tasks, comparing candidates' skills and knowledge with them (Pawełoszek 2017; Sanchez et al. 2018). In the present work, we use the lexical distance and the Dice's coefficient, to find the lexical similarity between knowledge topics and the topics of a thesaurus, where the combination of the measures allows handling the limitations generated by the length of topics when use only the editing distance (Kalmukov 2013).

With the use of thesauri and structural similarity measures, problems of ambiguity between the topics can be solved. For this, the taxonomy of the thesaurus is used to find common levels among them (Harispe et al. 2013), associating them to the same context, and in this case, to an area of knowledge. There are works where this concept is applied to align competencies, creating semantic networks (Malzahn et al. 2013; Sánchez et al. 2015) and graphs (Rácz et al. 2018). In others, they are used to determine suitable candidates to meet specific requirements, and align their profiles with job offers according to a thesaurus (Wordnet) (Montuschi et al. 2015). Also, the combination of structural similarity measures and thesauri allows the design of generic skills and accreditation requirements for university careers (Gluga et al. 2013) and smart learning environments (Paquette 2016). In the present work, a measure of similarity based on the adaptation of ant colony algorithms (ACO) is used (González-Eras and Aguilar 2015), to find semantic similarities, between the topics found in job advertisements and the taxonomies of the DISCO II and BLOOM thesauri, with the purpose of solving ambiguity problems. There are several works that use linked data vocabularies for the representation of professional offers or in the educational context (Smirnov et al. 2016; Faria et al. 2014; Sateli et al. 2017), but they have not been used for the representation of skills and knowledges of the competencies.

Regarding the calculation of the relevance of profiles according to the topics of the thesauri, Eq. 14 is a variation of the probabilistic model proposed in (Yuanhua and Zhailk 2011), sensitive to the frequency of topics aligned to thesauri and to the length of the documents, expressed in the topics (Definition 14). There are works where this model is used to identify profiles of authors according to characteristics (Weren et al. 2014); in others, the Okapi model is used for the recommendation where a weight is established to manage the frequency of the topics based on the characteristics of the collection (Nishioka et al. 2015). In our work, we consider the values of δ of 0.5 and 1 for the collections of academic profiles and job advertisements, respectively, because both collections present a frequency distribution and number of topics of the thesaurus used different. This weight represents the importance given to a topic in Eq. (16), according to its frequency and the length of the documents. With the relevance equation proposed in Definition 14, these differences can be handled to obtain results similar to those of other investigations.

In (Rosa et al. 2015), they present MultCComp, a multi-temporal context-aware system for competence management, which considers the workers' present and past contexts to help them to develop their competencies. Also, in (Rau 2017) is modeled the knowledge-component of the competencies of the students based on the hypothesis that knowledge-component models that describe the content knowledge and representational competencies should be more accurate than knowledge-component models that describe only content knowledge. They conclude that students can learn abstract content knowledge only if they have a prerequisite level of representational competencies, and that educational technologies should use adaptive knowledge-component models that capture representational competencies the student has not yet mastered. In our work, it is proposed the management of the competences via the implementation of similarity algorithms that make an alignment of the knowledge and skill topics found in academic profiles and job advertisements, against the topics present in a competence thesaurus. This establishes a semantic measure for the competence alignment between academic profiles and job advertisements, based on the similarity of their knowledge and skill topics.

There are *Applicant Tracking Systems*, like Jobscan, which analyze summaries and job descriptions, recognizing job titles, education levels and skills, to establish a ranking and to give recommendations. Also, there are *Automated Resume Screening* applications, such as Ideal, which select candidates according to their experiences, skills, among other things. In addition, there are patents based on techniques of natural language processing for analyzing candidates resumes (Dane 2012). These tools differ with our model, in terms of the profiles, the text processing techniques, and the knowledge bases, used to align the data. Our model uses linguistic patterns for the recognition and alignment of professional profiles based on competencies. In general, all these previous applications contribute to improve the candidate recruitment process, but none is oriented towards labor competence analysis to match academic profiles.

Conclusions and Future Work

The present work presents a model of alignment of academic profiles and job advertisements based on competencies, combining measures of lexical and semantic similarity, and the adaptation of a measure of relevance to the frequency of the topics and the length of the profiles. The obtained results allow determining the similarity of the profiles against the knowledge and skill topics of the DISCO II and BLOOM thesauri, and in this way to establish the relevance of the profile alignments based on a ranking measure.

The difference that our work presents with respect to others lies in the analysis of job advertisements in Spanish, through the combined use of lexical and semantic measures. The combination of lexical and semantic measures allows obtaining similar values on profiles in Spanish, in comparison with other works that analyze them in other languages. Another novelty is the use of the Okapi BM25 measure for the alignment of profiles, instead of the traditional TF-IDF algorithm, with a modification that makes this measure sensitive to the relationship between topic frequency and document length expressed in the topics. In addition, the use of two thesauri (DISCO II and BLOOM) allows the alignment process to be strengthened, using the topics contained in them. In this way, through the measures used, we have achieved the alignment of academic profiles and job advertisements, detecting the academic topics with which the job offers are most aligned, and giving feedback to those topics of job offers that did not align with any academic profile.

Our proposal can be applied in academic contexts, for the development of academic management systems and decision-making based on competencies, such as semantic search engines for educational resources, automatic creation of academic profiles of careers and subjects according to work requirements, evaluation of tasks, exams and courses based on professional competencies (González-Eras et al. 2017; Guevara et al. 2017; González-Eras and Aguilar 2019; Sánchez et al. 2015). Also, the model is applicable in the recruitment context, aligning the candidate resumes and job advertisements, according to their knowledge and skills; or for the automatic definition of staff training plans, according to the current competencies of the employees and the required skill and knowledge required in their job positions.

The flexibility of the proposed model allows it not only to be applicable to the context of Computer Science, but also to other areas of knowledge where the

knowledge bases in Spanish are available. In addition, our proposal can be used in other languages and domains, through the use of the appropriate NLP tools and knowledge bases, according to the required language and context, respectively. For instance, in our case, we have used Python and Stanford Core NLP libraries for Spanish as NLP tools (Manning et al. 2014). For the semantic analysis of the domain, we have used DISCO II, which is a multilingual thesaurus for different contexts (such as education, labor market, etc.) that offers the mapping of competencies in several languages.

The following steps are aimed at testing the schema in a corpus of larger profiles, as well as improving the detection mechanisms of topics in the profiles, as well as analyzing other groups of profiles and job offers. All of the above will allow us observing their alignment around the topics of the DISCO II and Bloom thesauri. Likewise, it is necessary to initiate experiments to replicate the proposed model with other knowledge bases in the domain of Computer Science, as is the case of ACM.

This work is part of an architecture for the analysis of job advertisements, which includes a phase of characterization of them according to competencies, representing linguistic and semantic aspects through descriptive and dialectical logic; which allows the recognition of the topics of knowledge and skill in the documents. Future work concentrates on completing the feedback phase, based on classification techniques and clustering of topics, and in the definition of learning analytic tasks (Sanchez et al. 2018) and intelligent recommender systems (Aguilar et al. 2017) of educational resources based on this architecture. Finally, future works will analyze the relationship between the *Applicant Tracking Systems*, the *Automated Resume Screening* applications, with our approach, in order to extend them with the labor competence analysis and academic profiles capabilities.

Data Availability The data is available in https://goo.gl/Q9d8Hx.

References

Aguilar, J., Valdiviezo, P., Cordero, J., Sánchez, M. (2015). Conceptual design of a smart classroom based on multiagent systems. In *Proceedings of Int. Conf. Artificial Intelligence* (471–477).

Aguilar, J., Valdiviezo, P., & Riofrio, G. (2017). A general framework for intelligent recommender systems. *Applied Computing and Informatics*, Elsevier, *13*(2), 147–160.

Alqadah, F., & Bhatnagar, R. (2011). Similarity measures in formal concept analysis. *Annals of Mathematics and Artificial Intelligence, 61*(3), 245–256.

Anderson, L. W., Krathwohl, D. R. and Bloom, B. S. (2001). "A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives", Allyn & Bacon.

Beckers, J. (2011). "Développer et évaluer des compétences à l'école: vers plus d'efficacité et d'équité". [Online]. Available: http://orbi.ulg.be/handle/2268/125331.

Blanco-González, J., Ortega-González, Y., et al. (2011). Ontological models for professional competences management. *Ingeniería Industrial, 32*(3), 224–230.

Dane, M. (2012). System and method for automatically processing candidate resumes and job specifications expressed in natural language into a normalized form using frequency analysis. U.S. Patent 8,117,024, issued February 14.

De Leenheer, P., Christiaens, S., & Meersman, R. (2010). Business semantics management: A case study for competency-centric HRM. *Computers in Industry, 61*(8), 760–775.

Dijkman, R., Dumas, M., Van Dongen, B., Käärik, R., & Mendling, J. (2011). Similarity of business process models: Metrics and evaluation. *Information Systems, 36*(2), 498–516.

Ehrig, M., Koschmider, A. and Oberweis, A. (2007). Measuring similarity between semantic business process models. Fourth Asia-Pacific conference on Conceptual Modelling, Australian Computer Society, Inc., pp. 71–80.

Faria, C., Serra, I., & Girardi, R. (2014). A domain-independent process for automatic ontology population from text. *Science of Computer Programming, 95*, 26–43.

Fazel-Zarandi, M. (2013). Representing and reasoning about skills and competencies over time. Ph.D. dissertation, Toronto Univ., Canada.

Gluga, R., Kay, J., & Lever, T. (2013). Foundations for modeling university curricula in terms of multiple learning goal sets. *IEEE Transactions on Learning Technologies, 6*, 25–37.

Gomaa, W. H. and Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*.

González-Eras, A. (2017). Caracterización de las competencias en los contextos laboral y académico en base a tecnologías semánticas. M.S. thesis. E.T.S.I.S.I., Universidad Politécnica de Madrid, Madrid, España.

González-Eras, A., & Aguilar, J. (2015). Semantic Architecture for the Analysis of the Academic and Occupational Profiles Based on Competencies. *Contemporary Engineering Sciences, 8*(33), 1551–1563.

González-Eras, A., Aguilar, J. (2019). Esquema para la actualización de Ontologías de Competencias en base al Procesamiento del Lenguaje Natural y la Minería Semántica. Revista Ibérica de Sistemas e Tecnologias de Informação, E17, (pp. 433–447).

González-Eras, A., Buendia, O., Aguilar, J., Cordero, J., Rodriguez, T. (2017). Competences as services in the autonomic cycles of learning analytic tasks for a smart classroom. In *Technologies and Innovation* (R. Valencia-García, et al., Eds.), Communications in Computer and Information Science Series, Vol. 749, Springer, pp. 211-226.

Guevara, C., Gonzalez, A., & Aguilar, J. (2017). The model of adaptive learning objects for virtual environments instanced by the competencies. *Advances in Science, Technology and Engineering Systems Journal, 2*(3), 345–355.

Harispe, S., Ranwez, S., Janaqi, S. and Montmain, J. (2013). Semantic measures for the comparison of units of language, concepts or instances, Parc scientifique, France: LGI2P/EMA Research Center.

Jones, K., Walker, S., & Robertson, S. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing & Management, 36*, 809–840.

Kalmukov, Y. (2013). Describing papers and reviewers' competences by taxonomy of keywords. *arXiv preprint*.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707–710.

Lundqvist, K., Baker, K., & Williams, S. (2011). Ontology supported competency system. *International Journal of Knowledge and Learning, 7*(3–4), 197–219.

Malzahn, N., Ziebarth, S., & Hoppe, H. (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. D., Raghavan, P., and Schütze, H. (2009). *An introduction to information retrieval*. Cambridge University Press.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit, In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, (pp. 55–60).

Mendonza, M., Perozo, N. and Aguilar, J. (2015). An approach for Multiple Combination of Ontologies based on the Ants Colony Optimization Algorithm. Asia-Pacific Conference on Computer Aided System Engineering, Quito: IEEE, Ecuador, pp. 140–145.

Montuschi, P., Lamberti, F., Gatteschi, V. and Demartini, C. (2015) A semantic recommender system for adaptive learning. *IT Professional*, 50–58.

Müller-Riedlhuber, H. (2009) The European dictionary of skills and competencies (DISCO): an Example of Usage Scenarios for Ontologies. I-SEMANTICS, pp. 467–479.

Müller-Riedlhuber, H. (2017). DISCO II the European dictionary of skills and competences. [Online]. Available: http://disco-tools.eu/disco2_portal/projectInformation.php

Nishioka, C., Große-Bölting, G. and Scherp, A. (2015). Influence of time on user profiling and recommending researchers in social media, 15th International Conference on Knowledge Technologies and Data-driven Business, ACM, pp. 9.

Ortiz Sánchez, C. L. (2016). Verificación de competencias académicas en base a niveles de habilidad mediante elementos semánticos, thesis, Dept. Computer Sciences and Electronics, Universidad Técnica Particular de Loja, Loja.

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

Paquette, G. (2016). Competency-based personalization process for smart learning environments. Learning, Design, and Technology. *International Compendium of Theory, Research, Practice, and Policy*, pp. 20–36.

Paquette, G., Rogozan, D. and Marino, O. (2012). Competency comparison relations for recommendation in technology enhanced learning scenarios, CEUR Workshop.

Pawełoszek, I. (2017). Ontological support for process-oriented competency management. In Information Technology for Management. Ongoing Research and Development, pp. 41–60.

Rácz, G., Sali, A., Schewe, K. D. (2018). Refining semantic matching for job recruitment: An application of formal concept analysis. In *International Symposium on Foundations of Information and Knowledge Systems*, pp. 322–339.

Rau, M. (2017). Do knowledge-component models need to incorporate representational competencies? *International Journal of Artificial Intelligence in Education, 27*, 298–319.

Reichhold, M., Kerschbaumer, J., Fliedl, G. and Winkler, C. (2012). Automatic generation of user role profiles for optimizing enterprise search. In 24th International Conference on Software & Systems Engineering and their applications, vol. 24, pp 241–248.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval, 3*(4), 333–389.

Rosa, J., Kich, M., & Brito, L. (2015). A multi-temporal context-aware system for competences management. *International Journal of Artificial Intelligence in Education, 25*, 455–492.

M. Sánchez, J. Aguilar, J. Cordero, P. Valdiviezo (2015).Basic features of a reflective middleware for intelligent learning environment in the cloud (IECL). *Asia-Pacific Conference on Computer Aided System Engineering*.

Sanchez, M., Cordero, J., Valdiviezo, P., Barba, L., & Chamba, L. (2018). Learning analytics tasks as services in smart classroom. *Universal Access in the Information Society Journal*, Springer, *17*(4), 693–709.

Sateli, B., Löffler, F., König-Ries, B., Witte, R. (2017). ScholarLens: extracting competences from research publications for the automatic generation of semantic user profiles. 3, *Peer J Computer Science*.

Smirnov, A., Kashevnik, A., Balandin, S., Baraniuc, O., Parfenov, V. (2016). Competency management system for technopark residents: Smart space-based approach. In *Internet of Things, Smart Spaces, and Next Generation Networks and Systems* (pp. 15–24). Springer, Cham.

Van Dongen, B., Dijkman, R., & Mendling, J. (2013). Measuring similarity between business process models. *Seminal Contributions to Information Systems Engineering*, 405–419.

Weren, E., Kauer, A. U., Mizusaki, L., Moreira, V. P., de Oliveira, J. and Wives, L. K. (2014). Examining multiple features for author profiling. *Journal of Information and Data Management*, pp. 266.

Worsley, M., & Blikstein, P. (2018). A multimodal analysis of making. *International Journal of Artificial Intelligence in Education, 28*, 385–419.

Yuanhua, L., & Zhailk, C. (2011). Lower-bounding term frequency normalization. In *Proc. 20th ACM international conference on Information and knowledge management*, pp. 7–16.