Applied Computer Science, vol. 16, no. 1, pp. 18–32
doi:10.23743/acs-2020-02

Submitted: 2020-03-13
Revised: 2020-03-20
Accepted: 2020-03-23

similarity metrics, recruitment, workers, college graduates

Mario BELLO[*], *Alejandra LUNA*[*], *Edmundo BONILLA*[*],
Crispin HERNANDEZ[*], *Blanca PEDROZA*[*], *Alberto PORTILLA*[**]

A NOVEL PROFILE'S SELECTION ALGORITHM USING AI

Abstract

In order to better understand the job requirements, recruitment processes, and hiring processes it is needed to know the people skills. For a recruiter this entails analyzing and comparing the curricula of each available candidate and determining the most appropriate candidate that the activities that are required by the position. This process must be carried in the shortest length of time possible. In this paper, an algorithm is proposed to identify those candidates, either workers or college graduates.

1. INTRODUCTION

The recruitment and selection is one of the most ancient areas of applied psychology, besides, it is one of the most important domains in talent management and human resources (Derous & Fruyt, 2016). The evolution of the labor market has caused the traditional recruitment process to not be enough. Nowadays, the internet has introduced new methods to carry the recruitment process, starting with a new form to generate or creating a resume and the way in which it is distributed to the companies (Kesler, Béchet, Roche, Torres-Moreno & El-Bèze, 2012). For this reason, organizations have started to use different technological platforms to lure personnel as part of the electronic recruitment (Esch & Mente, 2018). Some organizations have even started to adopt artificial intelligence methodologies in their recruitment processes (Esch, Black & Ferolie, 2019).

[*] Tecnológico Nacional de México/Instituto Tecnológico de Apizaco, 90300, Carretera Apizaco Tzompantepec, Esquina Av., Instituto Tecnológico S/N, Apizaco, Tlaxcala, México, edbonn@walla.co.il
[**] Smartsoft America Business Applications S.A. de C.V., 90806, Adolfo López Mateos S/N, Texcacoac, Chiautempan, Tlaxcala, México, aportilla@smartsoftamerica.com.mx

In order to help the recruiter in the search of the most suited profile, a profile selection algorithm is proposed in this paper. We use a methodology that employs two search criteria of profiles, guaranteeing that only the profiles that meet one of the criteria are analyzed. The profiles belong to real people. These were taken from web sites such as Indeed.com and Mexico's employment web page (www.empleo.gob.mx). At the same time, the methodology includes a pre-processing stage to standardize the profiles as well as eight similarity metrics (Cosine, Euclidean, Levenshtein, Dice, N-grams, Jaccard, Fuzzy distance and Q-grams), in charge of finding the similarity degree between the profile and the employment vacancy. This algorithm was developed for the platform I'm talenty, owned by Smartsoft America Business Applications. The purpose of the platform is early linking in which students, companies and educational institutions interact (I'm Talenty, 2019). Smartsoft is a TI company that develops innovative solutions for the Mexican market.

2. RELATED PAPERS

In the metric similarity field, an adjustable approach of object parameters to predict unknown data in soft incomplete fuzzy sets was proposed, this is based on the similarity metrics of fuzzy sets (Liu, Qin, Rao & Mahamadu, 2017). Kerzendorf (2019) presented the application of computational linguistic techniques into the literature within the field of astronomy, which is a result of the recommendation of scientific articles or reference texts. Another advance in the document grouping was originated with the implementation of the N-grams technique and the enhance squared cosine similarity (Bisandu, Prasad & Liman, 2018), the methodology consisted in preprocessing new scientifically articles. (Cheatham & Hitzler, 2013) It was focused on the ontological alignment systems that implement chain similarity metrics. A basic system was also developed to automatically select a similarity metric of each chain set for a pair of ontologies in execution time, based on the characteristics of the ontologies to be aligned.

In the work matching field, a deep neural network with imbedded layer model was implemented to predict the future professional details of an employee (Deng, Lei, Li & Lin, 2018), based on the data at an online resume. It was proved that this model was more effective and appropriated than methodologist such as Random Forest, XGB or a deep neural network. Another approach for the matching between the offer and the demand was proposed on ESCO ontology (Shakya & Paudel, 2019), which is a multilingual classification of abilities, competences, qualification, and European occupations. It uses the similarity score, which is a measure to show how alike two sets of data are. Another advance was done when a presentation of the approach for the alignment of online profiles and job announcements mixing themes of a thesauri with Levenshtein distance, the Dice's coefficient and the Okapi BM25 measure (González-Eras & Aguilar, 2019).

3. METHODOLOGY

In this work, we propose to use similarity metrics to analyze two elements and to find a similarity degree. The elements to be analyzed are:

- College graduate's profiles,
- Worker profiles,
- Job offers.

The proposed algorithm is illustrated in Figure 1, all the elements that comprise it are observed.

Fig. 1. Profiles selector algorithm

1. Data set – Is the type SQL database, where the profiles and the job offers are kept.
2. Standardize content – Before searching the similarity degree between the profiles and/or job offers the content of this must be standardized.
3. Similarity metrics – Once the elements to be analyzed have been standardized, the search for a similarity degree follows.
4. Results' assortment – Once all the elements had been analyzed either finding a college graduated for a job offer or searching for a worker profile, the result are ordered.
5. Ordered profiles – At last, it gives back the user the list of sorted profiles.

3.1. Data set

The profiles of college graduates and the job offers were obtained by using the Web Scraping technique on Indeed.com. The workers' profiles were obtained manually from Mexico's employment portal (www.empleo.gob.mx), due to the lower amount in this kind of profiles. We should keep in mind that the information gathered is in Spanish and matches real profiles and job offers.

Because of this, the extracted profiles from Indeed.com and Mexico's employment portal are not structured, for every user is free to describe his/her profile as he/she sees fit. Before saving the profiles in the database, it was necessary to sort them by type. The content of workers' profiles and college graduates was structured in fields such as:

- Id – It corresponds to an identification number and it was assigned in increasing order and without repetition.
- Career – It corresponds to the college graduate's career or the employment of the worker's profile. This field uses as a criteria of profile search.
- Specialty – It corresponds to the specialty that the candidate or worker have (in case it has it). This field is used as a profile search criterion.
- State – It indicates the state or city in which the workers or the college graduate lives.
- Description – It corresponds to the previous abilities knowledge and/or experience that are contained by the profile. This last field is taken into account to search for the similarity degree.

The content of job offers and workers profiles was structured in a similar way as the college profiles, this was because only one field was added:

- University – It corresponds to the university of the college graduate. For the worker's profile, it corresponds to their education level.

The database contains 154 records, divided on the college graduated profiles, the workers and the job offers.

3.2. Content standardization

Before looking for a candidate meeting a specific job offer, it is necessary to standardize the content of the elements under consideration.

This process implies the elimination of grammatical elements in the Spanish language. Some elements are:

- Specific or not specific articles – a/an/the,
- Possessives – Mine/your/his/ours,
- Demonstratives – This/that/these/those.

Other elements to eliminate are punctuation marks (dot, colon, semi-colon, quotation marks, etc.). The same rule applies for special characters (@, $, *, <, etc.).

At last, the rest of the content of the elements are switched to lower case. This is due to the possibility of a word being written in a different way. For instance, the programming language Java can be written as java, JAVA, etc. For a system engineer, this has the same meaning, however, for similarity metrics it implies a minimum difference. All the elements described previously are found in a dictionary, allowing you to continue adding more grammatical elements that can be discarded from the profiles.

The standardization process is carried in order to minimize orthographic mistakes besides eliminating those words that do not have useful information and interfere with the similarity metrics analysis. At the same time, this process is executed whenever it is necessary to look for a candidate for a work position, and it is applied to the profiles and offer.

3.3. Similarity metrics

To know the similarity among profiles and a job offer, we propose to use similarity metrics. A similarity metric reflects the closeness between two objects, it must correspond to the characteristics that are thought to be integrated in the data groups. In this document eight similarity metrics are used.

Euclidean distance. It is a standard metric for geometrical issues. It is the ordinary distance between two points, which can be easily measured with a ruler in a dimensional or tridimensional space. It is widely used in problems clustering, even in text clustering (Huang, 2018). The Euclidean distance of two documents is defined by equation (1).

$$D_E(\overrightarrow{t_1}, \overrightarrow{t_2}) = \left(\sum_{t=1}^{m} |W_{t,1} - W_{t,2}|^2\right)^{1/2} \quad (1)$$

Given two documents (t_1, t_2) represented by their vector terms $\overrightarrow{t_1}$ and $\overrightarrow{t_2}$, further it´s term sets $T = t_1, \ldots, t_m$.

Cosine similarity. This metric is based in angles and orientation between two vectors discarding their longitude, which means it is the same that the cosine of the angle between vectors (Sandhya, Lalitha, Govardhan & Anuradha, 2008). In equation (2), the cosine similarity is represented.

$$SIM_{c(\overrightarrow{t_1}, \overrightarrow{t_2})} = \frac{\overrightarrow{t_1} \times \overrightarrow{t_2}}{\|\overrightarrow{t_1}\| \|\overrightarrow{t_2}\|} \quad (2)$$

In the equation (2), $\overrightarrow{t_1}$ and $\overrightarrow{t_2}$ are considered to be m-dimensional vector on the terms set $T = t_1, \ldots, t_m$. Each dimension represents a term in the document, which is not negative. As a result, the value given to that metric is delimited in the interval [0, 1].

Jaccard's distance. It measures the similarity of two elements in a way that the intersection of the elements is divided between the data element union (Guo, Jerbi & O'Mahony, 2014). This metric is represented in equation (3).

$$SIM_J(\overrightarrow{t_1}, \overrightarrow{t_2}) = \frac{\overrightarrow{t_1} \cdot \overrightarrow{t_2}}{|\overrightarrow{t_1}|^2 + |\overrightarrow{t_2}|^2 - \overrightarrow{t_1} \cdot \overrightarrow{t_2}} \quad (3)$$

For the documents t_1 and t_2, the Jaccard coefficient compares the sum of the terms that appeared in any of the documents but that are not shared. The result of this metric is in the interval [0, 1].

Levenshtein distance. It is a proximity measurement between two strings that applies mainly to the sequence comparison in the linguistic domain, like detecting plagiarism and speech recognition (Behara, Bhaskar & Chung, 2018). Levenshtein distance calculates the less expensive set of intersections, eliminations, or substitutions that are required to transform a chain into another. This metric is represented in equation (4).

$$D_L(t_1, t_2) = min_s\left(\sum_{k=0}^{k=s} \beta_k\right) \quad (4)$$

It defines $S = S_0, S_1, \ldots, S_k, \ldots, S_s$ as the sequence of edition operations to transform the string t_1 to t_2, after the associated cost to each edition operation as $\beta_0, \beta_1, \ldots, \beta_k, \ldots, \beta_s$.

Fuzzy distance. These distances are used to compare different objects. Their definition is based in proximity, fuzzy set operation, etc. That makes different property prepositions in the similarity measures (Baccour, Alimi & John, 2014). The measurement based in the fuzzy union and intersection operations is defined on the equation (5).

$$M_{A,B} = \frac{\sum i(a_i \wedge b_i)}{\sum i(a_i \vee b_i)} \quad (5)$$

With the use of the equation (5), the similarity degree is obtained $M_{A,B}$ from fuzzy sets *A* y *B* (Pappis and Karacapilidis, 1993). Another metric of fuzzy similarity is the metric based on the difference and the sum of membership degrees. The equation (6) shows said metric.

$$S_{A,B} = 1 - \frac{\sum i|a_i - b_i|}{\sum i(a_i + b_i)} \quad (6)$$

In equation (6), the similarity degree $S_{A,B}$ is obtained from the fuzzy sets *A* y *B*.

Q-grams similarity. This metric divides a string into substrings of length q. The reason behind q-grams is that the characters sequence is more important than the character by themselves. The q-grams similarity is represented on equation (7).

$$SIM_Q(t_1, t_2) = 1 - \frac{\sum_{i=1}^{n} |match\ (q_i, Q_{t1}) - match(q_i, Q_{t2})|}{|Q_{t1}| + |Q_{t2}|} \quad (7)$$

The q-grams for a string t is obtained as a longitude vector space q over the string. We should also consider the longitude substrings q-1 and recognize the prefixes and suffixes of the string, called filler characters (#, %, $) are added at the beginning and at the end of the string (Gali, Mariescu-Istodor, Hostettler & Fränti, 2019).

Dice's similarity. This similarity metric is based on the absence and presence of words in two documents t_1 and t_2. Said metric is represented on equation (8).

$$SIM_d(t_1, t_2) = \frac{2|t_1 \cdot t_2|}{|t_1|^2 + |t_2|^2} \quad (8)$$

The main aspect of this metric is that it multiplies by two the total number of terms in two documents (Dice, 1945).

N-grams' similarity. It consists of the generalization of the longest common subsequence concept to include n-grams, by only including uni-grams (Kondrak, 2005). This metric is shown on equation (9).

$$SIM_n\left(\Gamma_{i,j}^{n}\right) = \frac{1}{n} \sum_{u=1}^{n} S_1(X_i + u, Y_i + u) \quad (9)$$

This metric formulates the similarity of n-grams as a function sn, where n is a fix parameter, while S_1 is equivalent to the function of the uni-grams similarity.

3.4. Sort the profiles and return to the user

When the similarity metrics finishes the evaluation of all the elements, what follows is to sort the results. The assortment of the results is done by using a Merge Sort algorithm, which carries a stable execution and stands out from other sorting algorithms (Sedgewick & Wayne, 2011). John Von Neumann developed said algorithm in 1945, and it is based on the divide and conquer technique. In broad sense the algorithm works as following:

- If the longitude of the list is 1 or 0, then it is sorted,
- The list is divided on two sorted lists of almost the same size,
- Each sub list is sorted repetitively by using the Merge Sort algorithm,
- The bub list that had been sorted are incorporated in one list.

The algorithm takes into account the mean of all the generated results by the similarity metrics of each analyzed element so that the candidate with the most similarity shows up at the beginning of the list, and the candidate with the less similarity, at the bottom. This list is finally sent back to the user for the decision-making.

4. TEST AND RESULTS

Fort the test of the proposed methodology, a SQL database was used, with a total of 154 registers among worker profiles, college graduates' profiles and job offers, taking into account that all the obtained data is from real individuals.

The search for both profiles can be done by using two criteria, by using the field Career or the field Specialty. For instance, if a worker is required, the search criteria will include what one or another field contains in the job offer, to later be able to search on the available profiles to proceed with the selection of profiles that matches the specific fields. The selected profiles are the analyzed using the metrics, in this way the qualification of all the profiles by the metrics is avoided. For the tests, there are two types of profiles to analyze, the profiles of workers and the profiles of college graduates, which are detailed below.

4.1. Workers' search

The process of searching workers is as follows, a driver vacancy with several abilities is taken as a base: "loading and unloading maneuvers, knowledge of traffic regulations, drive different transport units and 3 years' experience driving 3½ tons units". For this profile, the algorithm identifies eight candidates for the vacancy. In Table 1, only the best five profiles are shown, each profile is evaluated by the eight metrics generating eight results in a range [0,1]. If the value shown by the metrics is 1, it means that both the job offer and worker profile are identical, however if the value is 0, then they are different.

Tab. 1. Assessment of better qualified profiles for the driver vacancy

Profile Id	D_E	SIM_C	SIM_j	D_L	Fuzzy distance	SIM_Q	SIM_D	SIM_N
17	0.80	0.93	0.69	0.30	0.44	0.17	0.19	0.28
19	0.78	0.90	0.71	0.37	0.5	0.24	0.23	0.35
20	0.81	0.93	0.83	0.27	0.48	0.33	0.37	0.26
22	0.76	0.92	0.69	0.36	0.46	0.11	0.12	0.35
23	0.78	0.96	0.67	0.28	0.43	0.13	0.15	0.27

The algorithm determines that the No. 20 profile is the best candidate, this by obtaining the mean of the eight similarity metrics. In Figure 2, shows the similarity of each metric for the best profile found. It is worth mentioning that these results are the product of standardization of the content of elements.

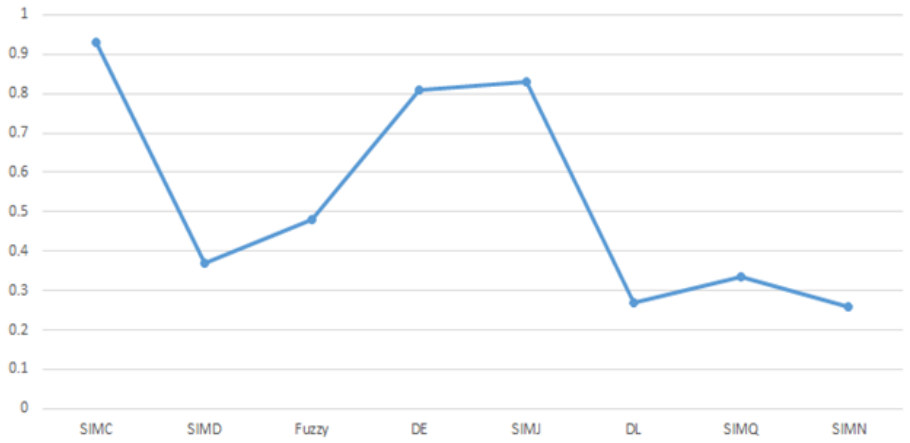

Fig. 2. Evaluation of each similarity metric for profile No. 20

Figure 3 shows how the best five profiles found are qualified by the similarity metrics.

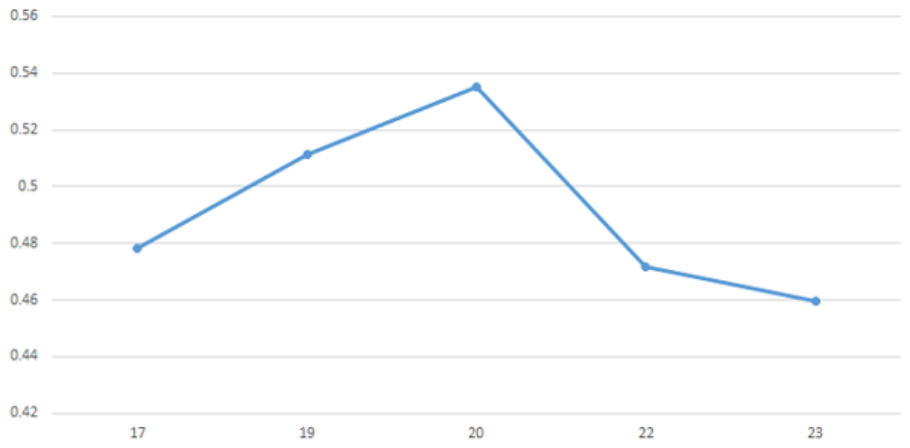

Fig. 3. Better qualified profiles for driver

It is confirmed in Figure 3, that the profile better qualified is profile No. 20, which has abilities such as: “1 year of experience driving 3 ½ ton units, current license, knowledge of traffic regulations, knowledge in the San Luis Potosí area

and its surroundings". The worker search process, standardization of the content elements, analysis of the elements by the similarity metrics and presentation of the candidates to the user were done in 3 seconds. In Figure 4, a small comparison of results it is shown when the contents of the elements had been standardized and when they had not been standardized and processed directly.

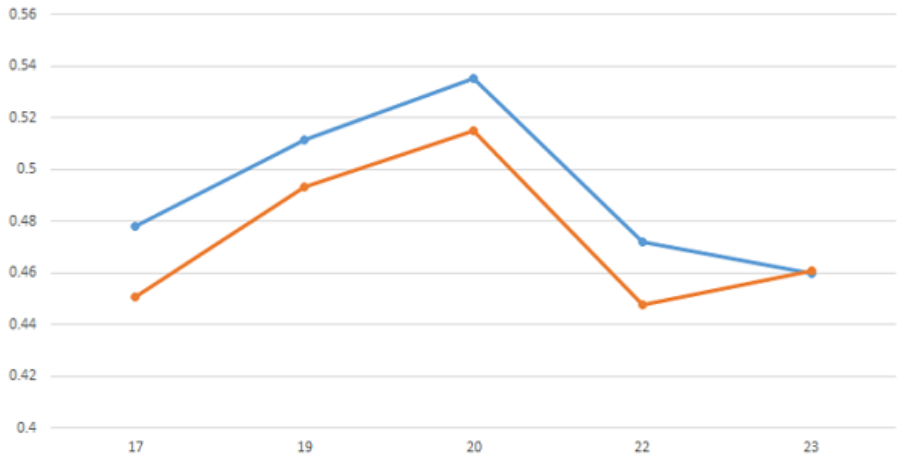

Fig. 4. Better five profiles result comparison with and without standardization

In Figure 4, the blue line belongs to the results after the standardization of the content, and the orange one belongs to the results obtained without standardization. Even though, in both cases, the best profile is profile No. 20. It is confirmed that by standardizing the similarity degree, it is higher. This is because the elements that are omitted on the profiles and job offers are more common, and this interferes with the metric analysis. In a way that these results improve when only the words that provide information or describe the person are taken into account.

4.2. College profiles search

The process of searching college graduated profiles is as follows, based on a job offer that requires a computer systems engineer with skills such as: "experience not required, availability to change residence, knowledge of object oriented programming with java, C#, Python, relational databases with MySQL, SQL server, Oracle or similar, basic knowledge of web design, HTML, CSS, JavaScript, English language skills". The algorithm identifies seven candidates. In Table 2, only shows the best five candidates. In a same way as in the search of a worker, if the value shown by metrics is 1, it means that the job offer and college graduate profile are identical, but if the value is 0, then they are different.

Tab. 2. Better qualified profiles for the computational systems offer

Profile Id	D_E	SIM_C	SIM_j	D_L	Fuzzy distance	SIM_Q	SIM_D	SIM_N
3	0.82	0.95	0.82	0.24	0.35	0.22	0.26	0.21
5	0.87	0.97	0.83	0.23	0.36	0.12	0.17	0.20
6	0.85	0.98	0.77	0.28	0.40	0.07	0.06	0.25
7	0.83	0.98	0.77	0.27	0.41	0.06	0.07	0.26
9	0.84	0.97	0.72	0.26	0.39	0.02	0.02	0.23

The algorithm determines that the best profile is No. 3. This is accomplished by obtaining the mean of all metrics. In Figure 5, the similarity degree for every metric is shown.

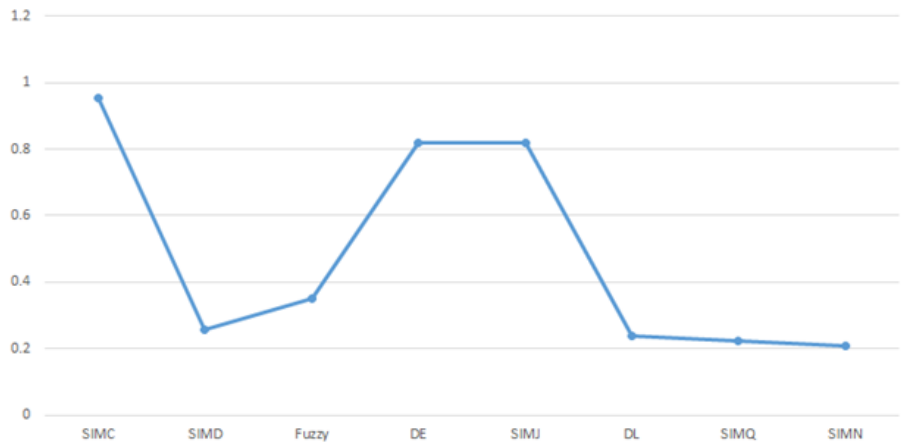

Fig. 5. Similarity metrics evaluation for profile No. 3

The Figure 6 shows the five best profiles evaluated by the similarity metrics.

It is confirmed that the best qualified profile by the metrics is No. 3, which has the following abilities: “knowledge about Microsoft Office, TOEIC certification, knowledge of databases with SQL, SQL server, MySQL, web page design with PHP and Java, maintenance of servers and knowledge in computer networks and 4 years’ experience”. The complete search of candidate process for this scenery was done in 3 seconds.

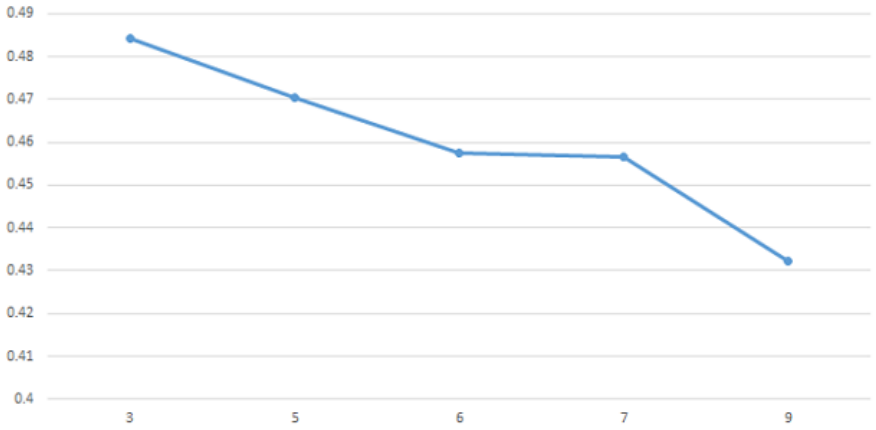

Fig. 6. Better qualified profiles for the computational systems engineer offer

In Figure 7, a comparative of the result when the process has been and has not been standardized is shown.

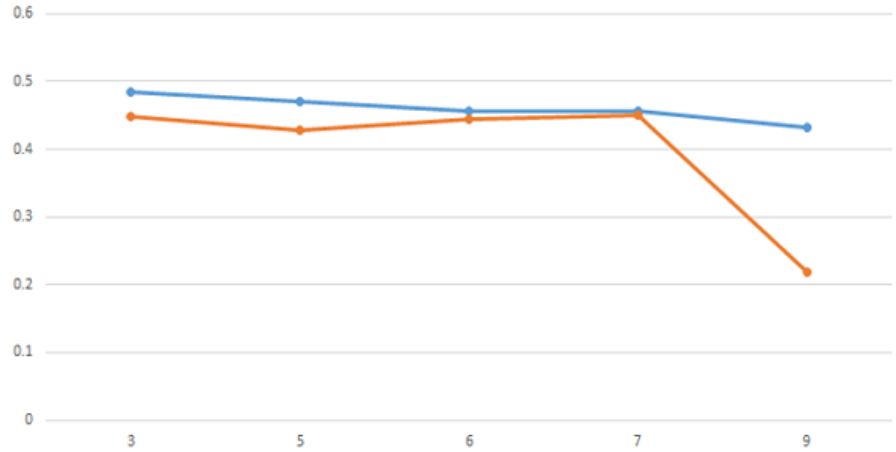

Fig. 7. Better five profiles result comparison with and without standardization

As the previous case, the blue line belongs to the result for the standardized profile, and the orange line belongs to the profile processing without standardizing. The same pattern, shown in the previous case, can be observed, and the similarity degree is higher when the elements have been standardized.

It also worth mentioning that the selection algorithm is not limited to the profile analysis in Spanish. It also works in English, however, a dictionary for that language has to be designed to be able to standardize the profiles. Figure 8 shows the evaluation of the algorithm for a robotics researcher vacancy. The analyzed profiles are in English.

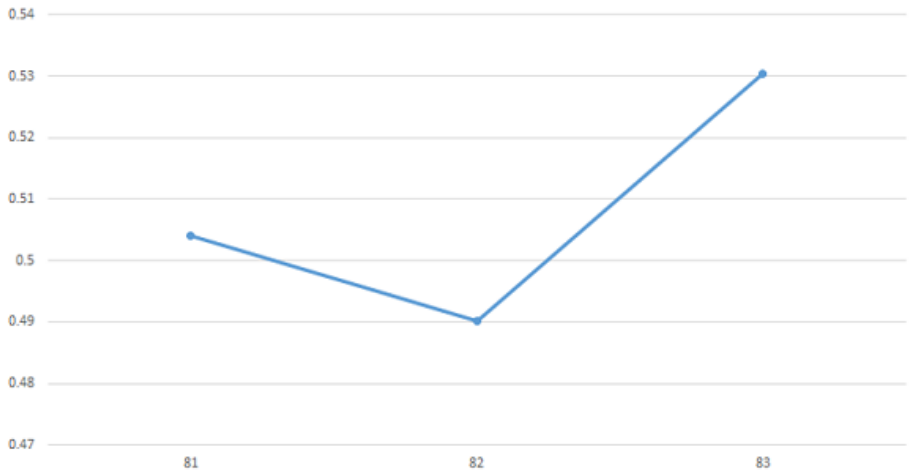

Fig. 8. English language evaluated profiles

The algorithm finds three candidates and the best evaluated for this is profile 83, which corresponds to a researcher.

The profile selection algorithm cannot be compared directly to other methodologies, because it carries different techniques in order to be able to align the profiles with job offers. Table 3 shows characteristics of the profile selector algorithm and the ESCO ontology (Shakya & Paudel, 2019).

Tab. 3. Characteristics of selector algorithm and ESCO ontology

Profile selection algorithm	ESCO ontology
Every time that a search is done, an standardization of the profiles is carried	Classifies abilities in the US job market
Takes abilities and knowledge that appear in the job offer into account	The abilities are defined and classified in the ontology
Two search criteria	Multicriteria
It conveys 8 similarity metrics	Uses “Similarity score” metric
Tested in Spanish and English	Functional for European languages

As it can be seen on Table 3, the methodologies have different characteristics. The strong points of the proposed methodology in this document is the standardization of the elements, the use of the eight metrics, and the sorting of the evaluated profiles. It can be seen that the processing prior to the analysis of similarity metrics is very important, since, when standardizing profiles, the degree

of similarity is greater, because only important information remains in the profile. At the same time, the eight metrics are taken into account for the evaluation and sorting of the profiles, this is in favor of having several experts that qualify in different ways when the evaluation is done and the sorting of the profiles that helps provide a quantitative perspective (by using the metric evaluation) on the found profiles.

5. CONCLUSIONS AND FUTURE WORK

According to our results, it can be said that it is possible to help a recruiter find the most suitable profile for a job offer, this is due to the reduction of a search range, allowing the recruiter to focus on the best evaluated profiles, besides, the search process is done in a few seconds, which means it is highly reduced, taking into consideration that a traditional recruitment method can take days or even weeks.

The disadvantage of this methodology is that it does not take into account the context of the profile. This means that a high degree similarity can be found in profiles within different areas. At the same time, this can be improved by implementing a semantic analyzer to process the elements before searching for the similarity degrees.

Future work focuses on the implementation of a stemming process. This process can be carried with the algorithm proposed by Porter (1980), the goal is to improve the similarity degree when looking for terms with a common root, this is because this kind of terms have similar meanings. Likewise, the content of the profiles can be segmented even more, which would allow a multicriteria search. At last, its functionality can be widened to other languages as long as the proper dictionary is built to be able to carry the standardization of the profiles.

REFERENCES

Baccour, L., Alimi, A., & John, R. (2014). Some notes on fuzzy similarity measures and application to classification of shapes, recognition of arabic sentences and mosaic. *IAENG International Journal of Computer Science*, *41*(2), 81–90.

Behara, K., Bhaskar, A., & Chung, E. (2018). Levenshtein distance for the structural comparison of od matrices. *40th Australasian Transport Research Forum (ATRF)*. Darwin.

Bisandu, D., Prasad, R., & Liman, M. (2018). Clustering news articles using efficient similarity measure and n-grams. *International Journal of Knowledge Engineering and Data Mining*, *5*(4), 333–348. doi:10.1504/IJKEDM.2018.095525

Cheatham, M., & Hitzler, P. (2013). String similarity metrics for ontology alignment. *International Semantic Web Conference*, *8219*, 294–309. doi: 10.1007/978-3-642-41338-419

Deng, Y., Lei, H., Li, X., & Lin, Y. (2018). An improved deep neural network model for job matching. *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 106-112. doi:10.1109/icaibd.2018.8396176

Derous, E., & Fruyt, F. D. (2016). Developments in Recruitment and Selection Research. *International Journal of Selection and Assessment*, *24*(1). doi:10.1111/ijsa.12123

Dice, L. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302. doi:10.2307/1932409

Esch, P., & Mente, M. (2018). Marketing video-enabled social media as part of your e-recruitment strategy: Stop trying to be trendy. *Journal of Retailing and Consumer Services*, *44*, 266–273. doi:10.1016/j.jretconser.2018.06.016

Esch, P., Black, J., & Ferolie, J. (2019). Marketing AI recruitment: The next phase in job application and selection. *Computers in Human Behavior*, *90*, 215-222. doi:10.1016/j.chb.2018.09.009

Gali, N., Mariescu-Istodor, R., Hostettler, D., & Fränti, P. (2019). Framework for syntactic string similarity measures. *Expert Systems with Applications*, *129*, 169–185. doi:10.1016/j.eswa.2019.03.048

González-Eras, A., & Aguilar, J. (2019). Determination of Professional Competencies Using an Alignment Algorithm of Academic Profiles and Job Advertisements Based on Competence Thesauri and Similarity Measures. *International Journal of Artificial Intelligence in Education*, *29*(4), 536–567.

Guo, X., Jerbi, H., & O'Mahony, M. (2014). An analysis framework for content-based job recommendation. In *International Conference on Case-Based Reasoning 2014*. Cork, Ireland.

Huang, A. (2008). Similarity measures for text document clustering. *New Zealand Computer Science Research Student Conference*, *6*, 49–56.

I´m Talenty (n.d.). *Intelligent platform for entailment student*. Retrieved January 10, 2019 from https://imtalenty.com/login.xhtml

Kerzendorf, W. (2019). Knowledge discovery through text-based similarity searches for astronomy literature. *Journal of Astrophysics and Astronomy*, *40*, 1–7. doi:10.1007/s12036-019-9590-5

Kessler, R., Béchet, N., Roche, M., Torres-Moreno, J., & El-Bèze, M. (2012). A hybrid approach to managing job offers and candidates. *Information Processing and Management*, *48*, 1124–1135. doi:10.1016/j.ipm.2012.03.002

Kondrak, G. (2005). N-gram similarity and distance. *String Processing and Information Retrieval*, *12*, 115–126. doi:10.1007/11575832_13

Liu, Y., Qin, K., Rao, C., & Mahamadu, M. (2017). Object-parameter approaches to predicting unknown data in an incomplete fuzzy soft set. *International Journal of Applied Mathematics and Computer Science*, *27*(1), 157–167. doi:10.1515/amcs-2017-0011

Pappis, C., & Karacapilidis, N. (1993). A comparative assessment of measures of similarity of fuzzy values. *Fuzzy Sets and Systems*, *56*(2), 171–174. doi:10.1016/0165-0114(93)90141-4

Porter, M. (1980). An algorithm for suffix stripping. *Program*, *40*, 211–218.

Sandhya, N., Lalitha, Y., Govardhan, A., & Anuradha, K. (2008). Analysis of similarity measures for text clustering. *Computer Science Journals*, *2*(4), 1–10.

Sedgewick, R., & Wayne, K. (2011). *Algorithms*, 4th Edition (pp. 244–336). Princenton.

Shakya, A., & Paudel, S. (2019). Job-Candidate Matching using ESCO Ontology. *Journal of the Institute of Engineering*, *15*(1), 1–13.