How can we measure the similarity between résumés of selected candidates for a job?

Conference Paper · July 2015

CITATIONS
6

READS
2,917

4 authors:

Luis Adrián Cabrera-Diego
Jus Mundi
32 PUBLICATIONS **131** CITATIONS
SEE PROFILE

Barthélémy Durette
Solstice Scop SA
26 PUBLICATIONS **377** CITATIONS
SEE PROFILE

Juan-Manuel Torres-Moreno
Université d´Avignon et des Pays du Vaucluse
237 PUBLICATIONS **1,573** CITATIONS
SEE PROFILE

Marc El-Bèze
Université d´Avignon et des Pays du Vaucluse
124 PUBLICATIONS **853** CITATIONS
SEE PROFILE

How Can We Measure the Similarity Between Résumés of Selected Candidates for a Job?

Luis Adrián Cabrera-Diego[1,2], Barthélémy Durette[2], Matthieu Lafon[2], Juan-Manuel Torres-Moreno[1,3] and Marc El-Bèze[1]
[1]LIA, Université d'Avignon et des Pays de Vaucluse, Avignon, France
[2]Adoc Talent Management, Paris, France
[3]École Polytechnique de Montréal, Montréal, Canada

Abstract— *Several researches in Natural Language Processing (NLP) have developed e-Recruitment systems. Despite these researches, none of them has been interested in the way the similarity and distance measures, using different vector weights, behave when they have to determine the likeness of résumés. Therefore, in this paper we present a comparative analysis of 5 measures using different vector weights done over a large set of French résumés. The aim is to know how these measures behave and whether they validate the idea that selected résumés have more in common with themselves than with the rejected résumés. We make use of NLP techniques and ANOVAs to do the comparative analysis. The results show that the selection of measures and vector weights must not be considered negligible in e-Recruitment projects, specially in those where the résumés' likeness is measured. Otherwise, the results may not be reliable or with the expected performance.*

Keywords: e-Recruitment, résumé analysis, similarity measures, matching systems, data mining

1. Introduction

During the last 15 years, the massification of computers and the Internet have had an impact on the way humans search for jobs and employees [1], [2], [3]. Internet has become the main medium to select and recruit candidates [4]. The use of information and communication technologies to recruit and select candidates for a job offer is what has been called *e-Recruitment* [4], [5], [6].

E-Recruitment has brought several benefits to Human Resources Managers (HRM), employers and job seekers. Nowadays, employers reach larger audiences [7], [8], HRM reduce their operating costs [7] while job seekers can search easily a job offer [9]. Although the e-Recruitment has brought the aforementioned benefits, some undesirable consequences have also arisen for HRM: an important increment in the number of unqualified applications [10] and the recruiters' difficulty to manage correctly and rapidly the great amount of received data [11], [12].

Many researches, usually in Natural Language Processing (NLP), have developed systems in order to increase the performance of HRM. These systems can be classified in three types: systems that extract specific résumé[1] data [14], [15], [16], systems that extend the information of job offers and/or résumés [14], [16] and systems that try to find the best candidate(s) for a job offer using ontology matching [17], semantic similarity [6], [8], automatic learning [18], [19] or relevance feedback [3]. Nevertheless, even if the aim of these researches is to create tools to assist HRM, to our knowledge, none of them have analyzed the role that similarity and distance measures, with different vectors weights, play in the likeness calculation of résumés. Furthermore, these researches have been developed and tested mainly using small datasets.

For these reasons, we present in this paper a comparative analysis of 5 measures applied with at least 3 different vector weights. The aim of this paper is to determine experimentally how the likeness of résumés behaves using these measures and vector weights; the results may be of help to determine in the future the best methodology to create e-Recruitment tools. The analysis is done over a large set of French résumés organized by job offer and which has been used and annotated by expert recruiters. We make use of NLP techniques in order to preprocess the data (i.e. language and résumé detection), and of statistical tests of Analysis of Variance (ANOVA) to asses each hypothesis.

The structure of this paper is the following: first, in Section 2, we present the data used in this project and their preprocessing. Then, in Section 3, we describe the experimental method. Later, we present and discuss the results in Section 4 and Section 5, respectively. Finally, in Section 6, we present the conclusions and future work.

2. Data

The corpus used in this paper comes from a HRM firm and is a collection of résumés, motivation and recommendation letters, diplomas, interview minutes and social networks invitations (LinkedIn, Twitter, Facebook, among others). The corpus is organized by job positions, which in turn are divided into candidates. It is annotated with meta-data that

[1]In some researches and books it is possible to find the Latin locution *curriculum vitæ (CV)* instead of the term *résumé*. However, for [13] both expressions are synonyms. Therefore, in this paper we will use the résumé as common term.

allow us to determine, for each job applicant, its unique ID, the applied position, the receiving date of each application and the last recruitment phase reached by the applicant (*Analyzed*, *Contacted*, *Interviewed* or *Hired*). French is the main language in the corpus although it is possible to find documents in English, Spanish and German. Table 1 shows the number of files, job offers, and job applications in the corpus.

Table 1: Number of job offers, job applications and files in the corpus.

Job offers	Job applications	Files
296	29,368	47,388

The four recruitment phases were classified into two classes: *Selected* and *Rejected*. The first class, corresponding to the phases *Contacted*, *Interviewed* or *Hired*, represents the candidates that are approached by a recruiter. The second class, contains the candidates that are only *Analyzed* but not contacted after reading their résumé. Figure 1 presents the histogram of Selected candidates, measured by percentage, in the corpus; the median is 40.94% and the mean $42.93\% \pm 1.44$.

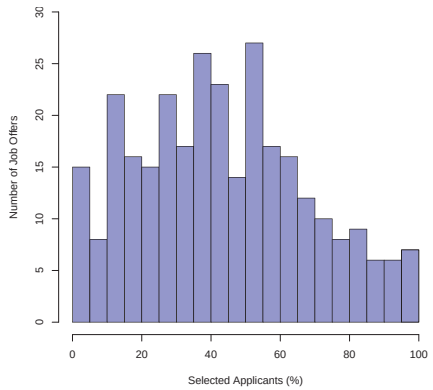

Fig. 1: Percentage of Selected candidates by the number of job offers.

2.1 Document Conversion and Language Recognition

Four types of documents are analyzed in this work: PDF (*.pdf*), Microsoft Word (*.doc* and *.docx*), OpenDocument Text (*.odt*) and Rich Format Text (*.rtf*). They represent 80.52% (38,161 files) of the corpus; the rest belongs mainly to HTML (social networks invitations) and image files. To be able to apply NLP techniques we first converted these files into plain UTF-8 text. For this we used:

- *Calibre Ebook Management*[2] for files having a *.pdf*, *.docx*, *.odt* or *.rtf* extension. The accentuated letters of PDF files are verified to know if they were correctly coded (see [20] for a discussion).
- *Catdoc*[3] is used only for files with *.doc* extension.

In order to detect only the French documents we used the Google's *Compact-Language-Detector (CLD2)*[4] through its Perl module[5]. The CLD2 is a tool that makes use of probabilities and 4-grams of letters to predict the language of documents[6]. In the corpus, 32,845 documents are in French (69.31% of the total corpus and 86.06% of the analyzed file formats).

2.2 Résumé Detection

To sort out the résumés from other types of documents, like motivation letters, interview minutes and publications lists, we developed a Résumé Detector based on a Support Vector Machine (SVM) [21] through LIBSVM [22].

2.2.1 Training

The training corpus was established through a manual classification of résumés (699) and other documents (635), all in French, from a collection of spontaneous applications[7]. We tested 2 different SVM kernels (linear and radial) following the procedure proposed by [23]. They were tuned up through a grid-search and a five-fold cross-validation.[8] Table 2 presents the results of the cross validation of the SVM using the best parameters for the linear and radial kernel. These results are presented in terms of precision, recall and F-score.

Table 2: Parameters, precision, recall and F-score for the linear and radial kernel.

Linear ($C = 0$)						
Subcorpora	L_1	L_2	L_3	L_4	L_5	Mean[9]
Precision	0.972	1.00	0.950	0.971	0.971	0.979
Recall	0.986	0.971	0.971	0.992	0.992	0.982
F-score	0.979	**0.985**	0.960	**0.982**	**0.978**	**0.977**
Radial ($C = 0$; $\gamma = 1 \times 10^{-4}$)						
Subcorpora	L_1	L_2	L_3	L_4	L_5	Mean[10]
Precision	0.979	0.963	0.964	0.951	0.932	0.952
Recall	0.986	0.949	0.971	0.985	0.992	0.977
F-score	**0.982**	0.956	**0.967**	0.968	0.961	0.964

[2]http://calibre-ebook.com/

[3]http://site.n.ml.org/info/catdoc/

[4]https://code.google.com/p/cld2/

[5]"Lingua::Identify::CLD" https://github.com/ambs/Lingua-Identify-CLD

[6]Documentation available at: https://code.google.com/p/cld2/wiki

[7]Job applications that are not related to any job offer and in consequence they do not belong to the experimental corpus.

[8]For the different models, all the documents from the training subcorpora passed through a basic preprocessing: stopwords suppression and stemming (Porter's Algorithm).

[9]Mean F-score is obtained from the average Precision and Recall.

[10]Idem.

As seen in Table 2, the best results are obtained with the use of the linear kernel, with an average F-score of 0.977; this performance was expected, as the number of features (in this case words) is much greater than the 2 possible classes (Résumés and Other documents) [23]. The Résumé Detector was implemented using a SVM with a linear kernel and the complete training corpus.

2.2.2 Evaluation

The evaluation corpus is a multilingual and heterogeneous set of 240 documents (résumés, motivation letters, publications lists, diplomas, etc.), divided into 4 groups of 60 documents: *French Résumés*, *Résumés in other languages*, *Other French documents* and *Other documents in other language*. It was generated manually by a non-expert recruiter who classified documents randomly chosen from the corpus. Two expert recruiters were asked separately to classify the files from the evaluation corpus into the same groups. The agreement between both expert recruiters was calculated with *Cohen's Kappa* ($\kappa = 93\% \pm 0.04$) and *Kendall's W* ($W = 0.905$ $p\text{-}value = 2.58 \times 10^{-13}$). In order to evaluate the Résumé Detector, each evaluation corpus (*Recruiter 1* and *Recruiter 2*) passed though the document conversion and language detection. Then, the Résumé Detector was utilized to determine which French documents, from both processed corpora, were résumés. Table 3 shows the results from this evaluation in terms of Precision, Recall and F-score; a mean for each measures is presented as well.

Table 3: Evaluation of the Résumé Detector in terms of Precision, Recall and F-score.

Corpus	Precision	Recall	F-score
Recruiter 1	0.964	0.916	0.939
Recruiter 2	**0.982**	**0.965**	**0.973**
Mean	0.973	0.940	0.956

The Résumé Detector reaches a good performance over the evaluation corpora with an average F-score of 0.956. Some cases where the Résumé Detector and the recruiters did not agree are the documents which are bilingual résumés or motivation letters that have short résumé attached.

We applied the Résumé Detector over the documents of the corpus that were detected previously in French. From this task the module detected 22,439 French résumés (47.35% of the total corpus and 68.31% of the converted French documents).

2.3 Résumé Uniqueness

We found that in the corpus there are candidates which have more than one résumé for the same job offer. This happens either because the applicants have sent several résumés for one application or because the applications have been forwarded more than once. The information inside the multiple résumés may or not be exactly the same.

To avoid false or biased results from these cases, we validated the résumé uniqueness in each job offer. The validation is done using 3 tests applied sequentially over all the possible couples of résumés in a job offer[11]:

1) One résumé by candidate: both résumés must come from two different applicants.
2) Résumés with different content: the Linux tool *Diff*[12] is used to validate if both résumés are equal[13].
3) Not equal e-mails: e-mails addresses[14] from the two résumés must be different.

After the verification, for each existing problematic couple, the oldest résumé, according to the receiving date, is deleted.[15] Nevertheless, if a *Rejected* résumé is identical to a *Selected* one, the former will be the deleted one. This exception only applies when a candidate has sent two applications to the same job offer and the first one was *Selected* and the second one, in consequence, *Rejected*.

3. Methodology

We inferred that the résumés of *Selected* candidates are more alike with themselves than with the rest of résumés sent to a job offer. This is because the candidates with résumés fulfilling the characteristics of a job offer are the only contacted by a recruiter. In this paper, we would like to know how the use of certain measures and data weighting affects this inference.

If we consider a *Likeness Score* (LS) as the measure where the higher the value the more alike are the résumés and a set J as all the possible couples of résumés for a job offer, our inference can be represented mathematically with Equation 1.

$$\overline{LS}(J_S^c) < \overline{LS}(J_S) \tag{1}$$

where $\overline{LS}$ is the average Likeness Score, J_S is the subset of J that contains all the possible couples of *Selected* résumés and J_S^c is the complement of J_S. Figure 2 shows an example of possible couples of subset J_S and J_S^c with 3 *Selected* résumés and 3 *Rejected* ones.

3.1 Data Representation

To use a Vector Space Model (VSM) [24] as data representation, we converted each résumé into 3 different vectors. These vectors are constructed from unigrams, bigrams and skip bigrams (SU4) [25], [26] of words. It should be noted

[11]The number of possible couples for a certain job offer is given by all the possible combinations of two résumés (C_2^n) taken from the total number of résumés (n).

[12]http://www.gnu.org/software/diffutils/

[13]The *Diff* tool has been configured to ignore the multiple white spaces and lines but also to be case-insensitive.

[14]The e-mails were detected using a regular expression.

[15]Since the tests are applied in pairs, one résumé can be deleted due to several reasons.

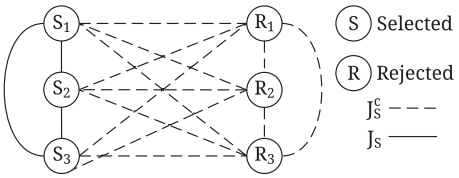

Fig. 2: Example of the possibles couples of the subset J_S and J_S^c with 3 *Selected* candidates and 3 *Rejected* ones.

that before the résumés' n-grams extraction, we lowercased all documents. Also, we removed all the punctuations marks, numbers and stop-words[16] from each document. And we reduced the documents' lexicon through Porter's algorithm for French (stemming).[17] These tasks were done to reduce the possible noise in the text, the size of the VSM and the curse of dimensionality [27].

In addition, the 3 resulting vectors have been represented by 3 types of weights: *absolute frequency*, *relative frequency* and *TF-IDF*. In the case of the TF-IDF, the values are calculated with respect to each job offer. The relative frequencies are obtained résumé by résumé.

3.2 Similarity and Distance Measures

To calculate the Likeness Score of résumés, we implemented 3 similarity measures (*Cosine Similarity*, *Jaccard's Index*, *Dice's Coefficient*) and 2 distance measures (*Euclidean Distance*, *Manhattan Distance*). Table 4 recalls the formula of each measure.

Table 4: Formulæ of the similarity and distance measures.

Measure	Formula	Measure	Formula
Cosine Similarity	$\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$	Jaccard's Index	$\frac{\lvert X \cap Y \rvert}{\lvert X \cup Y \rvert}$
Manhattan Distance	$\sum \lvert x_i - y_i \rvert$	Dice's Coefficient	$2\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$
Euclidean Distance	$\sqrt{\sum \lvert x_i - y_i \rvert^2}$		

Each measure was applied by type of n-grams (unigrams, bigrams and skip-bigrams) and type of weight (absolute frequency, relative frequency or TF-IDF values). In the case of Jaccard's Index and Dice's Coefficient, we only make use of the absolute frequency as weight. The reason is that we implemented their binary version, which only takes into account the existence or absence of elements. As well, we only applied Cosine Similarity to absolute frequency and TF-IDF values as the results using absolute or relative frequencies will be always the same [28, Page 111].

In order to have only one Likeness Score by type of weight, we decided to merge the 3 n-grams' Likeness Scores into one by a simple combination. This combination consists in multiplying each Likeness Score by an *influence factor* and making the addition of the resulting values. The influence factor of the 3 types of n-grams has been settle to $1/3$, giving the same leverage to all of them. However, the influence factor can be changed independently on the condition that the sum of them is equal to one. This merge is quite *naïve* but our purpose is to follow an *a fortiori* principle. If this combination method leads us to good results, the use of more sophisticated merging methods or influence factor setting may lead us to better results.

The calculation of each Likeness Score was parallelized using GNU Parallel [29].

3.3 Statistical Test

To know whether the Likeness Scores of the groups, J_S and J_S^c are statistically different, we performed a two-tailed Analysis of Variance (ANOVA) for independent groups.

Owing to the characteristics of the corpus, not all the job offers are analyzable with our methodology. We found that there are 63 job offers where it does not exist at least one French résumés in J_S or J_S^c, making impossible to calculate any measure. These cases represent the job offers without résumés from a Selected or Rejected applicant, see Figure 1, and the job offers ($\approx$ 25) where non-French résumés are dominant. We found as well 9 cases where the number of résumés of group J_S prevent us to verify whether the measure distribution is normal. These 72 job offers were deleted from the analysis.

Before doing any ANOVA we suppressed the outliers from the groups J and J_S^c of each analyzable job offer (224). We defined the outliers as the values that are 1.5 times the interquartile range ($IQR = Q_3 - Q_1$) below the first quartile (Q_1) or above the third quartile (Q_3) [30, Page 208]. After deleting the outliers, we verify that both groups fulfill the following two assumptions, which are necessary to do an ANOVA:

- Normal distribution: a Shapiro-Wilk Test ($\alpha = 0.05$) is applied to verify data normality. As this test can only be used in groups that contain between 3 and 5,000 elements [31], the groups with less than 3 elements are not considered as Gaussian. The groups with more than 5,000 elements are considered as normal even if it may be a violation of the assumption. However, the ANOVA is a robust test where the normality assumption can be discarded with minor effects [32, Page 424].
- Variance equality: The homogeneity of variances is tested with a Bartlett's Test ($\alpha = 0.05$). In case of heterogeneous variances, the ANOVA is only done if the biggest variance is not greater than 4 times the smallest one [32, Page 354].

In case one of the groups of a job offer does not surpass the previous conditions, the job offer is considered not analyzable (NA).

[16] List taken from the Perl's module "Lingua::StopWords".

[17] We used the Perl's module "Lingua::Stem::Snowball", an interface for the stemmers of the Snowball project (http://snowball.tartarus.org/).

Once the ANOVA of a job offer has been calculated, we consider that the averages of the Likeness Score for both groups, $\overline{LS}(J_S)$ and $\overline{LS}(J_S^c)$, are statistically different only if the ANOVA's $p\text{-}value < 0.05$.

4. Results

The results for Cosine Similarity, Manhattan Distance, Euclidean Distance, Dice's Coefficient and Jaccard's Index are shown in Table 5. The outcomes for the first 3 measures are divided by vector component weight: absolute frequency, relative frequency and TF-IDF.

For all the results, we present the number of job offers where the average Likeness Score ($\overline{LS}$) for both groups is statistically different ($p\text{-}value < 0.05$) and statistically equal ($p\text{-}value \geq 0.05$). As well, we present the number of cases that did not surpass the ANOVA's conditions (NA), the number of job offers where the elements of J_S have more in common with themselves (J_S+) and the cases where the elements of J_S^c have more in common with themselves (J_S-). The total number of analyzable job offers in the corpus was 224.

Table 5: Results for Cosine Similarity, Manhattan distance, Euclidean distance, Dice's Coefficient and Jaccard's Index.

		$p\text{-}value$			
		< 0.05		≥ 0.05	NA
		J_S+	J_S-		
Cosine Similarity	Absolute/Relative Frequency	163	11	44	6
	TF-IDF	158	10	55	1
Manhattan distance	Absolute Frequency	53	90	63	18
	Relative Frequency	164	7	50	3
	TF-IDF	59	86	62	17
Euclidean distance	Absolute Frequency	69	83	58	14
	Relative Frequency	124	30	62	8
	TF-IDF	69	78	61	16
Jaccard's Index		164	1	58	1
Dice's Coefficient		164	2	56	2

As seen in Table 5, there are seven cases that clearly follow our inferred behavior (Cosine Similarity; Manhattan and Euclidean Distances with relative frequencies; Jaccard's Index and Dice's Coefficient) and 4 cases where it is clear that the inference does not behave as inferred (Euclidean Distance with absolute frequency and TF-IDF; Manhattan Distance with TF-IDF and absolute frequency).

With the purpose of comparing easily the results between the different measures and vectors' weights, 3 rates and one score have been established:

$$SR = \frac{\text{Total}_{\text{Significant}}}{\text{Total}_{\text{Analyzable job offers}}} \quad (2)$$

$$TR = \frac{\text{Total}_{\text{Significant}} + \text{Total}_{\text{Not Significant}}}{\text{Total}_{\text{Analyzable job offers}}} \quad (3)$$

$$IR = \frac{\text{Total}_{J_S+}}{\text{Total}_{\text{Statiscally different}}} \quad (4)$$

$$RS = \sqrt[3]{SR * TR * IR} \quad (5)$$

The *Significant rate* (Equation 2) indicates the ratio between the number of job offers with a significant ANOVA test and the total number of analyzable job offers in the corpus. The *Testing rate* (Equation 3) expresses the proportion of job offers tested with an ANOVA regarding the total number of analyzable job offers. Our inference about the résumés' Likeness Scores ($J_S^c < J_S$) is measured with the *Inference rate* (Equation 4), which is the number of job offers following the expected behavior per the number of job offers with an ANOVA $p\text{-}value < 0.05$. The *Ranking Score* (Equation 5) is a value which allow us to rank the measures according to their Significant, Testing and Inference rates. For the three rates and the score the higher the value, the better (the maximum value is 1 while the minimum is zero). Table 6 shows the values of the 3 rates and the score for each measure.

Table 6: Significance rate (SR), Testing rate (TR), Inference rate (IR) and Ranking Score (RS) for each measure.

		SR	TR	IR	RS
Cosine Similarity	Absolute/Relative Frequency	**0.776**	0.973	0.936	0.890
	TF-IDF	0.750	**0.995**	0.940	0.888
Manhattan distance	Absolute Frequency	0.638	0.919	0.370	0.600
	Relative Frequency	0.763	0.986	0.959	0.896
	TF-IDF	0.647	0.924	0.406	0.623
Euclidean distance	Absolute Frequency	0.678	0.937	0.453	0.660
	Relative Frequency	0.687	0.964	0.805	0.810
	TF-IDF	0.656	0.928	0.469	0.658
Jaccard's Index		0.736	**0.995**	**0.993**	**0.899**
Dice's Coefficient		0.741	0.991	0.987	0.898

Taking into account the results presented in Table 6, we can see that in terms of the Significant rate, the highest score is the one of Cosine Similarity using frequencies (0.776); in terms of Testing rate, the highest rates are obtained by Cosine Similarity with TF-IDF and Jaccard's Index (0.995). Regarding the Inference Rate, the highest score is for Jaccard's Index (0.993). And with respect to the Ranking Score the leading one (0.889) is also for Jaccard's Index. The overall lowest scores are those obtained by Manhattan Distance using absolute frequency with a Significant Rate of 0.638, a Testing Rate of 0.919, an Inference rate of 0.370 and Ranking Score of 0.600.

Finally, from the results of the 11 analysis we can point out three points:

- Seven analysis clearly present results that follow the expected behavior.
- Relative frequencies as vector weight improve the performance of distances measures.
- Binary measures have comparable performance to Cosine Similarity.

5. Discussion

The performance of the Manhattan and Euclidean distances are not the expected one. We attended to have similar results, for all types of weights, like Cosine Similarity. However, only the use of relative frequencies, in both measures, give the expected outcome. Moreover, the performance of Manhattan distance can be greatly improved, passing from the worst Ranking Score (0.600) to the third best (0.896), better than Cosine Similarity using frequencies.

It can be thought that the disagreement behavior obtained by Manhattan and Euclidean distances using TF-IDF or absolute frequency is linked to the Gaussian assumption. We considered as normal the groups J_S^c or J_S having more than 5,000 elements, as their size exceed the superior limit of the Shapiro-Wilk Test. Nonetheless, the effect of this decision may not be relevant if it is taken into account that only 45 cases (20.08%) of the analyzable job offers (224) have at least one group with more than 5,000 elements. Moreover, 44 of these cases have always a homogeneous variance and one of them, depending on the measure and the vector weight, can have a homogeneous variance or not.

Actually, we think that the disagreement behavior is related to the intrinsic characteristics of Manhattan and Euclidean distances. Unlike Cosine similarity, Dice's Index and Jaccard's Coefficient, which are measures always delimited by the interval of values $[0, 1]$, Manhattan and Euclidean distances are measures that can have a $[0, \infty)$ interval. This means that the superior interval limit is not defined and that it is restricted to the size and lexical richness of the documents. Therefore, two measures of the same distance may have different interval limits and comparing them may not be equivalent. For example, for two completely different documents, their distance X means 0% in common, while for two documents half different, their distance X means 50% in common; both distances have the same value X but different scale, making their comparison incompatible.

In our case, the résumés are not limited neither in size nor in vocabulary, hence the use of not normalized versions of Manhattan and Euclidean distances, i.e. with a closed interval, are not reliable in most cases. The exception is Manhattan Distance with relative frequencies, in this case this type of weight works like a distance normalization as it close the interval[18] into $[0, 2]$. It may be thought that the use of Euclidean Distance with relatives frequencies would be an exception as well, however, the relative frequency does not close the interval.

In order to understand better our results, we analyzed the job offers marked as NA. We found that all the NA cases are job offers with a heterogeneous variance. This means, that all the analyzed job offers have more than 3 elements and those between 3 and 5,000 elements have a normal distribution. In addition, we can see that the number of cases with heterogeneous variances arises when we make use of distance measures without relative frequencies.

The performance of Cosine Similarity with TF-IDF did not turn out as anticipated. We assumed that the use of TF-IDF would boost the performance of Cosine Similarity, as the components of the vectors would be weighted by their importance [33]. Nevertheless, the difference of the Ranking Score between the use of frequencies and TF-IDF values is about -0.225%; for the others rates the difference are: Significant Rate -3.35%, Testing Rate $+2.26\%$ and Inference Rate $+0.42\%$.

Finally, the performance of Dice's Index and Jaccard's Coefficient exceeded our expectation. We never thought that only the presence or absence of n-grams could be enough to determine the inferred behavior; however, we found that binary measures are sensible enough to determine résumés' likeness. Thus, we can infer that Selected résumés have a specific vocabulary which is not present in the Rejected résumés. Moreover, this means that the frequency of "terms" is not relevant for recruiters, instead of it, the most important thing is the appearance or not of "terms" related to the job offer requirements.

6. Conclusions and Future Work

In this paper, we made a comparative analysis of 3 similarity measures (Cosine Similarity, Dice's Coefficient, Jaccard's Index) and 2 distance measures (Euclidean and Manhattan distances). All the measures, excepting Dice's Coefficient, Jaccard's Index, were compared with at least 3 types of vector weights (absolute frequency, relative frequency and TF-IDF values). The objective was to determine how the use of different measures and vector weight affects the likeness detection of Selected résumés, i.e. résumés from applicants contacted by a recruiter.

This work was done over a large annotated recruitment corpus coming from a HRM firm. We made use of NLP techniques in order to detect the French résumés from the corpus. As well, we utilized an Analysis of Variance (ANOVA) to determine how the 5 measures considered the likeness of the résumés. And therefore, to compare with our inference: whether Selected résumés have more in common with themselves than the rest of résumés do.

Results varied according to type of vector weight and to measure. The use of Manhattan or Euclidean Distances may not be reliable to measure the likeness of résumés if some

[18] If two vectors X and Y using relative frequencies are completely different, their Manhattan Distance becomes $\sum X_i + \sum Y_i = 1 + 1 = 2$. Therefore, the maximum value possible in this case is 2.

considerations are not taken. The document size and lexical richness affects strongly these measures. Therefore, it may be better to use their normalized versions.

Cosine Similarity has shown the best results when it is used with frequencies, relative or absolute. Nonetheless, their performance was reduced when we used TF-IDF.

The use of Jaccard's Index and Dice's Coefficient, presented a good performance and exceeded our expectations. Moreover, we think that the use of these measures to find likeness between résumés, for example in matching systems, may give good results.

And we believe, according to the results obtained from Jaccard's Index and Dice's Coefficient, that there must be a specific vocabulary that could lead us to detect easily the résumés from candidates that should be Selected by a recruiter or not.

As future work, we are going to analyze other languages, like English, in order to determine whether our methodology can be defined as language independent. In addition, we will implement new procedures to reduce the lexicon, like lemmatizers or other stemmers. As well, we will analyze how other types of vector weight affect the tests, for example other TF-IDF methods. We will improve the method utilized to merge the likeness score of n-grams, for example using a weighted mean, calculating the influence factor of each type of n-gram in the likeness score or creating one vector with all the n-grams. New distances will be also tested, like normalized versions of Manhattan and Euclidean distances, or non-binary versions of Jaccard's Index and Dice's Coefficient. The use of non parametric ANOVAs and/or robust ANOVAs is expected. Finally, we will do more inferences about the Rejected résumés, and about the job offer and the résumés.

Acknowledgments

This project has received the support from the Agence National de la Recherche et de la Technologie (ANRT) from the CIFRE convention 2012/0293b and from the Consejo Nacional de Ciencia y Tecnología (CONACyT) grant 327165.

The authors would like to thank Amandine Bugnicourt, Clémence Chardon and Elodie Chevalier from Adoc Talent Management. They helped us to understand better the HRM tasks and gave us several ideas in order to develop some tools used in this project.

References

[1] R. Rafter, K. Bradley, and B. Smyth, "Automated Collaborative Filtering Applications for Online Recruitment Services," in *Adaptive Hypermedia and Adaptive Web-Based Systems*, ser. Lecture Notes in Computer Science, P. Brusilovsky, O. Stock, and C. Strapparava, Eds. Springer Berlin Heidelberg, 2000, vol. 1892, pp. 363–368.

[2] P. De Meo, G. Quattrone, G. Terracina, and D. Ursino, "An XML-Based Multiagent System for Supporting Online Recruitment Services," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 37, no. 4, pp. 464–480, July 2007.

[3] R. Kessler, N. Béchet, M. Roche, J.-M. Torres-Moreno, and M. El-Bèze, "A hybrid approach to managing job offers and candidates," *Information Processing & Management*, vol. 48, no. 6, pp. 1124–1135, 2012.

[4] C. Bizer, R. Heese, M. Mochol, R. Oldakowski, R. Tolksdorf, and R. Eckstein, "The impact of semantic web technologies on job recruitment processes," in *Wirtschaftsinformatik 2005*. Springer, 2005, pp. 1367–1381.

[5] V. Radevski and F. Trichet, "Ontology-based systems dedicated to human resources management: An application in e-recruitment," in *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, ser. Lecture Notes in Computer Science, R. Meersman, Z. Tari, and P. Herrero, Eds. Springer Berlin Heidelberg, 2006, vol. 4278, pp. 1068–1077.

[6] L. Yahiaoui, Z. Boufaïda, and Y. Prié, "Semantic Annotation of Documents Applied to E-Recruitment." in *Proceedings of SWAP 2006, the 3rd Italian Semantic Web Workshop*, 2006, pp. 1–6.

[7] D. S. Chapman and J. Webster, "The use of technologies in the recruiting, screening, and selection processes for job candidates," *International Journal of Selection and Assessment*, vol. 11, no. 2-3, pp. 113–120, 2003.

[8] P. Montuschi, V. Gatteschi, F. Lamberti, A. Sanna, and C. Demartini, "Job recruitment and job seeking processes: how technology can help," *IT Professional*, vol. 16, no. 5, pp. 41–49, 2014.

[9] D. Looser, H. Ma, and K.-D. Schewe, "Using formal concept analysis for ontology maintenance in human resource recruitment," in *Proceedings of the Ninth Asia-Pacific Conference on Conceptual Modelling-Volume 143*. Australian Computer Society, Inc., 2013, pp. 61–68.

[10] E. Faliagka, L. Kozanidis, S. Stamou, A. Tsakalidis, and G. Tzimas, "A personality mining system for automated applicant ranking in online recruitment systems," in *Web Engineering*. Springer, 2011, pp. 379–382.

[11] R. Rafter, B. Smyth, and K. Bradley, "Inferring relevance feedback from server logs: A case study in online recruitment," in *The 11th Irish Conference on Artificial Intelligence and Cognitive Science (AICS 2000)*, 2000.

[12] F. Trichet, M. Bourse, M. Leclère, and E. Morin, "Human resource management and semantic web technologies," in *Information and Communication Technologies: From Theory to Applications, 2004. Proceedings. 2004 International Conference on*. IEEE, 2004, pp. 641–642.

[13] M. A. Thompson, *The global resume and CV guide*. Chichester, New York: Wiley, 2000.

[14] A. Singh, C. Rose, K. Visweswariah, V. Chenthamarakshan, and N. Kambhatla, "PROSPECT: a system for screening candidates for recruitment," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 659–668.

[15] W. B. A. Karaa and N. Mhimdi, "Using ontology for resume annotation," *International Journal of Metadata, Semantics and Ontologies*, vol. 6, no. 3, pp. 166–174, 2011.

[16] D. Çelik, A. Karakas, G. Bal, C. Gultunca, A. Elçi, B. Buluz, and M. C. Alevli, "Towards an Information Extraction System Based on Ontology to Match Resumes and Jobs," in *Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual*. IEEE, 2013, pp. 333–338.

[17] V. Senthil Kumaran and A. Sankar, "Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping (EXPERT)," *International Journal of Metadata, Semantics and Ontologies*, vol. 8, no. 1, pp. 56–64, 2013.

[18] E. Faliagka, L. Iliadis, I. Karydis, M. Rigou, S. Sioutas, A. Tsakalidis, and G. Tzimas, "On-line consistent ranking on e-recruitment: seeking the truth behind a well-formed CV," *Artificial Intelligence Review*, pp. 1–14, 2013.

[19] R. Kessler, J. M. Torres-Moreno, and M. El-Bèze, "E-gen: Profilage automatique de candidatures," *TALN 2008, Avignon, France*, pp. 370–379, 2008.

[20] L. A. Cabrera-Diego, J.-M. Torres-Moreno, and M. El-Bèze, "SegCV : traitement efficace de CV avec analyse et correction d'erreurs," in *Actes de TALN 2013*, 2013, pp. 707–714.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
[22] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
[23] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Department of Computer Science, National Taiwan University, Tech. Rep., 2003.
[24] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.
[25] X. Huang, F. Alleva, H.-w. Hon, M.-y. Hwang, and R. Rosenfeld, "The SPHINX-II Speech Recognition System: An Overview," *Computer, Speech and Language*, vol. 7, pp. 137–148, 1992.
[26] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Association for Computational Linguistics, 2004, pp. 74–81.
[27] R. E. Bellman, *Adaptive control processes: a guided tour*. Princeton, New Jersey: Princeton University Press, 1961.
[28] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008.
[29] O. Tange, "GNU Parallel - The Command-Line Power Tool," *;login: The USENIX Magazine*, vol. 36, no. 1, pp. 42–47, Feb. 2011.
[30] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, 5th ed. Hoboken, New Jersey: John Wiley & Sons, 2010.
[31] P. Royston, "Remark AS R94: A remark on algorithm AS 181: The W-test for normality," *Applied Statistics*, pp. 547–551, 1995.
[32] D. Howell, *Fundamental statistics for the behavioral sciences*, 8th ed. Wadsworth, California: Cengage Learning, 2013.
[33] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.