Job Recommendation based on Job Seeker Skills: An Empirical Study

Jorge Valverde-Rebaza Ricardo Puma Paul Bustios Nathalia C. Silva

Department of Scientific Research, Visibilia, CEP 13560-647, São Carlos, SP, Brazil
{jvalverr, rpuma, pbustios, ncsilva}@visibilia.net.br

Abstract.

In the last years, job recommender systems have become popular since they successfully reduce information overload by generating personalized job suggestions. Although in the literature exists a variety of techniques and strategies used as part of job recommender systems, most of them fail to recommending job vacancies that fit properly to the job seekers profiles. Thus, the contributions of this work are threefold, we: i) made publicly available a new dataset formed by a set of job seekers profiles and a set of job vacancies collected from different job search engine sites; ii) put forward the proposal of a framework for job recommendation based on professional skills of job seekers; and iii) carried out an evaluation to quantify empirically the recommendation abilities of two state-of-the-art methods, considering different configurations, within the proposed framework. We thus present a general panorama of job recommendation task aiming to facilitate research and real-world application design regarding this important issue.

Keywords: Job matching, job seeking, job search, job recommender systems, person-job fit, LinkedIn, word embedding.

1 Introduction

Nowadays, job search is a task commonly done on the Internet using job search engine sites like LinkedIn[1], Indeed[2], and others. Commonly, a job seeker has two ways to search a job using these sites: 1) doing a query based on keywords related to the job vacancy that he/she is looking for, or 2) creating and/or updating a professional profile containing data related to his/her education, professional experience, professional skills and other, and receive personalized job recommendations based on this data. Sites providing support to the former case are more popular and have a simpler structure; however, their recommendations are less accurate than those of the sites using profile data.

Personalized job recommendation sites implemented a variety of types of recommender systems, such as content-based filtering, collaborative filtering, knowledge-based and hybrid approaches [AlO12]. Moreover, most

In: A. Jorge, R. Campos, A. Jatowt, S. Nunes (eds.): Proceedings of the Text2StoryIR'18 Workshop, Grenoble, France, 26-March-2018, published at http://ceur-ws.org

[1] https://www.linkedin.com

[2] https://www.indeed.com

of these job recommender systems perform their suggestions based on the full profile of job seekers as well as by considering other data sources such as social networking activities, web search history, etc. Despite the fact that many data sources can be useful to improve the job recommendation, previous studies showed that the best person-job fit is possible when the personal skills of a job seeker match with the requirements of a job offer [Den15].

Based on the person-job fit premise, we propose a framework for job recommendation based on professional skills of job seekers. We automatically extracted the skills from the job seeker profiles using a variety of text processing techniques. Therefore, we perform the job recommendation using TF-IDF and four different configurations of Word2vec over a dataset of job seeker profiles and job vacancies collected by us. Our experimental results show the performances of the evaluated methods and configurations and can be used as a guide to choose the most suitable method and configuration for job recommendation.

The remainder of this paper is organized as follows. In Section 2, we briefly describe the natural language processing methods we are used in our experimental setup. In Section 3 we present our proposal, including a new dataset collected by us and the framework for job recommendation. In Section 4, we show our experimental results. Finally, in Section 5, we offer conclusions and directions for future work.

2 Background

In this section, we briefly describe two methods used in our experiments: *Term Frequency-Inverse Document Frequency* (TF-IDF) and *Word2vec*. Moreover, for Word2Vec we also present two models commonly used over it: *Continuous Bag-of-Words* (CBOW) and *Skip-gram*.

2.1 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF assigns weights to the words as a statistical measure used to evaluate the relevance of a word in document of a corpus [Sal88]. This relevance is proportional to the number of times a word appears in the document and inversely proportional to the frequency of the word in the corpus.

This method has been successful in topic identification over large text datasets, but its performance decrease when applied over small ones as those commonly found in job descriptions. However, TF-IDF has been applied to deal with recommendation obtaining interesting results [Dia13] [Dia14].

2.2 Word2vec

Word2vec is a general predictive model for learning vector representations of words [Mik13b]. These vector representations, also called *word embeddings*, capture distributional semantics and co-occurrence statistics [Mik13a]. There are two Word2vec models we can use to obtain word embeddings: CBOW and Skip-gram.

Continuous Bag-of-Words model (CBOW). This model predicts a target word based on the n words before and n words after the target word [Mik13b]. For example, in the following sentence:

Lorem ipsum dolor sit amet

CBOW will predict the word *dolor* taking as inputs $n = 2$ words before and after it, i.e. *Lorem*, *ipsum*, *sit* and *amet*. These words are called the context of the target word and their quantity is a parameter of the model.

Skip-gram. Rather than predicting a word based on its context, Skip-gram aims to predict the context based on one word [Mik13a]. For instance, based on our previous example, skip-gram will try to predict the words *Lorem*, *ipsum*, *sit* and *amet* having only the word *dolor* as input.

3 Proposal

In this section, we describe our framework for job recommendation. We narrow down the scope and focus on recommendation of job vacancies for Information Technology (IT) professionals acting in the Brazilian market. The proposed framework, depicted in Fig.1, is composed by three stages: data collection, data preparation and recommendation.

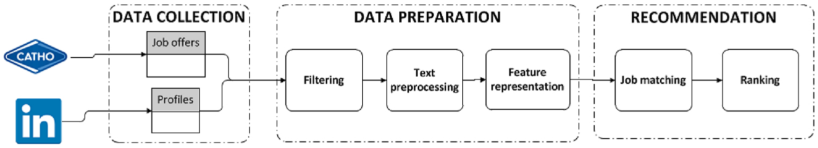

Fig. 1. Pipeline of the recommendation framework.

3.1 Data collection

We automatically collected a set of job vacancies/offers from the Brazilian recruitment site called *Catho*[3] and a set of Brazilian professional profiles from the well-known LinkedIn. We make available these datasets in a public repository[4] with personal data anonymised. It is important to note that we collected more data from similar sites but, due to the validation issues, we only managed to work with these two sources in our framework.

To perform job offers scraping, we created a list of keywords from the IT industry and used them as search terms. For each keyword, we search all the related job offers using Catho's search engine and save the retrieved results in our database; thus, the content's quality is highly related to the quality of the Catho's search engine. Additionally, the scraper is set up to avoid duplicate job offers, thus all the job offers are unique. On the other hand, to perform professional profiles scrapping, we created a list of areas of professional practice from the IT industry and, from that, we search among the professional contacts of first, second and third degree of our research group using Linkedin's search engine and save the retrieved results in our database; thus, all the professional profiles also are unique.

We use text mining approaches to process both profiles and job offers data. Therefore, we selected the work experience, education and competencies/skills from the profiles and, the description and title from the job offers. Finally, we concatenate these fields into a new one and discard the original fields, thus we end up with a document-like representation for each job offer and professional profile.

3.2 Data preparation

Although we retrieved data from job search sites using only IT keywords, there were still some job offers that do not correspond to this field, then, the first step in this phase is filtering out job offers that do not belong to the IT field. To achieve this, we use a dictionary of weighted IT terms to match each job offer in its document-like format. This way, we calculate the weighted sum of the appearances of each word of the job offer in the dictionary and divided it by the appearances of the rest of words in the document (job offer). Finally, we get a score with a value from 0 to 1, where a higher value indicates that the offer contains many relevant words on IT and it is very likely that corresponds to this field. Subsequently, we select only those job offers with a value of this score greater than 0.5. This setback only happens with the job offers since profiles were collected only into a IT professionals network.

Once job offers and profiles are filtered, the second step is text preprocessing. In this task, we perform stop words removal, tokenization and lemmatization for the Portuguese language.

The third step, feature representation, aims to represent these documents (job offers and profiles) as vector space models. For this purpose, we adopted two approaches: word embeddings and TF-IDF. The latter technique does not require so much effort to be implemented unlike the former. From the variety of word embedding representations we selected Word2Vec, which has different variants. We explore the two model architectures CBOW and Skip-Gram, and also the use of n-grams (bigrams and trigrams) in order to find the variation that best fit our problem. This way, we tested 5 different representations, TF-IDF, Word2Vec using CBOW, Word2Vec using Skip-Gram, Word2Vec using CBOW with n-grams and Word2Vec using Skip-Gram with n-grams. For the Word2vec models, a vector space size of 200 was selected after some initial experimentation.

For both word embedding and TF-IDF representation, we only used the corpus composed by the job offers. Although we lose some data, it was necessary since we realized that job seeker profiles added some noise because of the existence of professionals with a very different background and skill set looking for a job on IT, which could foster spurious relations among skills. Finally, we transform both job offers and profiles into these 5 new representations and then proceed to use them in the recommendation phase. In Table 1, we can see the description of the corpora used for our word embeddings.

[3] `https://www.catho.com.br`

[4] http://visibilia.net.br/text2story-job-recomendation/

Table 1. Word embeddings description

Dataset	# Documents	# Tokens
Profiles	50	111970
Job offers	3877	157576

3.3 Recommendation

In this last phase, given a certain profile with a proper representation, we select a group of the nearest job offers based on the distance to that profile (job matching). In the case of TF-IDF representation, we use the cosine distance while for word embeddings, we use the relatively new Word Mover's Distance (WMD) [Kus15]. Once retrieved the top "k" job offers for the profile, we sort them in descending order based on the inverse of this distance (ranking).

4 Experimental Results

In this section, we present extensive empirical experiments focused on evaluating the quality of job recommendations. For these experiments, we take the case of recommending a set of job offers given a specific professional profile.

Our data set is composed by 50 professional profiles from LinkedIn and 3877 job offers from Catho. Both profiles and job offers correspond to Brazilian professionals and companies from the IT field. Due to the extensive of the IT field, professional profiles can also differ a little bit among them. Table 2 shows the distribution of subfields within our sample of 50 professional profiles which reflects the greater number of developers and BI consultants.

Table 2. Professional profiles breakdown

Subfield	Profiles
Architect	5
BI consultant	10
Developer	24
Manager	2
Technical Support	9

First, we use our framework to generate 10 job offer recommendations for 50 different profiles. Thus, for each evaluated technique, we obtained a total of 500 recommendations. Second, a group of 5 Resource Human professionals evaluated manually these recommendations and allocate a score ranging from 1 to 10. The more accurate or suitable the recommendation, the greater the score. In order to make the results more understandable, we standardize these scores dividing them by the maximum score. Third, once these scores are obtained, we averaged them and also calculated Precision and Minimum Effectiveness (ME).

Precision for a single profile by dividing the number of relevant documents (recommendations with a score greater than 0.5) by all the retrieved documents (total of recommendations); then, we average this precision over all the profiles. On the other hand, the Minimum Effectiveness (ME) allocates a score of 1 if at least one out of the 10 recommendations for a profile has a score greater or equal to 0.5, otherwise it allocates 0. Thus, we average this value to have an estimator of the global effectiveness of the system of 10 job recommendations per profile. In Table 3, we show the result of applying these metrics over our dataset for the 5 different evaluated techniques.

Table 3. Results of job offers recommendation for each technique used. Each column shows the average of the metric over all the recommendations.

	Score	**Precision**	**ME**
TF-IDF	0.588	0.775	**0.96**
Word2Vec-CBOW	0.548	0.765	0.92
Word2Vec-SkipGram	**0.590**	**0.814**	**0.96**
Word2Vec-ngrams-SkipGram	0.582	0.784	0.92
Word2vec-ngrams-CBOW	0.580	0.783	**0.96**

Here, we can observe that Word2Vec with Skip-Gram obtains a slightly better average score than TF-IDF, which has the second best average. On the other hand, Word2Vec with Skip-Gram clearly gets a better average precision over all the other techniques by a good margin and it is the best option according to the three metrics. The Word2Vec variant using Skip-Gram with n-grams ranked second. Furthermore, we also observe that not all the profiles were given a good recommendation as the maximum value of the average minimum effectiveness is 0.96 (48 out of 50 profiles). This last metric is highly dependent on the quality of the filtering process and the variety of job offers since there can be a shortage of offers for some specific profiles. Finally, we can see that the two versions of Word2Vec using n-grams perform better than the Word2Vec with CBOW according to all metrics used. Plus, these n-gram versions have a slightly better average precision than TF-IDF, but a lower average score.

5 Conclusion

In this paper, we proposed a framework for job recommendation task. This framework facilitates the understanding of job recommendation process as well as it allows the use of a variety of text processing and recommendation methods according to the preferences of the job recommender system designer. Moreover, we also contribute making publicly available a new dataset containing job seekers profiles and job vacancies.

Future directions of our work will focus on performing a more exhaustive evaluation considering a greater amount of methods and data as well as a comprehensive evaluation of the impact of each professional skill of a job seeker on the received job recommendation.

Acknowledgments

This work was partially supported by the São Paulo Research Foundation (FAPESP) grants: 2016/08183-5, 2017/14995-5, 2017/15070-5, 2017/15247-2 and 2017/17312-6.

References

[AlO12] Shaha T Al-Otaibi and Mourad Ykhlef. "A survey of job recommender systems". In: *International Journal of the Physical Sciences* 7.29 (2012), pp. 5127–5142. ISSN: 19921950. DOI: 10.5897/IJPS12.482.

[Den15] N Deniz, A Noyan, and O G Ertosun. "Linking Person-job Fit to Job Stress: The Mediating Effect of Perceived Person-organization Fit". In: *Procedia - Social and Behavioral Sciences* 207 (2015), pp. 369–376.

[Dia13] M Diaby, E Viennet, and T Launay. "Toward the next generation of recruitment tools: An online social network-based job recommender system". In: *Proc. of the 2013 IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining, ASONAM 2013* (2013), pp. 821–828. DOI: 10.1145/2492517.2500266.

[Dia14] M Diaby and E Viennet. "Taxonomy-based job recommender systems on Facebook and LinkedIn profiles". In: *Proc. of Int. Conf. on Research Challenges in Information Science* (2014), pp. 1–6. ISSN: 21511357. DOI: 10.1109/RCIS.2014.6861048.

[Kus15] M Kusner et al. "From word embeddings to document distances". In: *Proc. of the 32nd Int. Conf. on Machine Learning, ICML'15*. 2015, pp. 957–966.

[Mik13a] T Mikolov et al. "Distributed Representations of Words and Phrases and Their Compositionality". In: *Proc. of the 26th Int. Conf. on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada, 2013, pp. 3111–3119. URL: http://dl.acm.org/citation.cfm?id=2999792.2999959.

[Mik13b] T Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[Sal88] G Salton and C Buckley. "Term-weighting approaches in automatic text retrieval". In: *Information Processing and Management* 24.5 (1988), pp. 513–523. ISSN: 0306-4573. DOI: https://doi.org/10.1016/0306-4573(88)90021-0. URL: http://www.sciencedirect.com/science/article/pii/0306457388900210.