Improving Employee Recruitment Through Data Mining

Visar Shehu(✉) and Adrian Besimi

South East European University, 1200 Tetovo, Macedonia
{v.shehu,a.besimi}@seeu.edu.mk

Abstract. Companies have always struggled with recruiting suitable candidates. In this age of data, we believe that the process of recruiting candidates is broken. This paper presents our efforts to improve the process by introducing data analytics and smart decision making. Recruiters and recruiting companies can benefit from such findings by analyzing key performance indicators and recommendation systems when recruiting new candidates. Furthermore, we propose an approach of identifying employment trends as well as new skills that are required by the job market. The procedure is fully automatic and relies on machine learning approaches.

Keywords: Recruitment · Clustering · Data analytics · Prediction systems

1 Introduction

Data mining is the process or methodology of extracting previously unknown information from large datasets. Often the age we are living nowadays has been referred to as the age of data. This, of course, is due to the huge amount of data being generated. As a consequence, there is a need for complex and trustworthy means to analyze this data with the purpose of generating actionable knowledge.

Companies are in a constant struggle to hire and keep employees. With the commodity provided by web technologies, many are relying on publishing vacancies online and then continuing with the hiring process offline. This offline process is consisted mainly of interviewing candidates, which can be extended to multiple phases. While there is data generated from this process, one cannot expect to have a standardized approach to the recruitment process.

Recruiting and headhunting agencies are trying to benefit from online vacancy publishing platforms to ease the process of identifying and hiring suitable candidates. Among the most popular websites of this nature are LinkedIn, Monster, Jobs.ch, cvmanager.ch, etc. They all use various approaches and are targeting diverse audiences, but they all share a common thing: the ability to generate big amounts of data.

Many researchers have focused on using data mining or artificial intelligence applications to identify and hiring suitable candidates. Jantan et al. in [5] propose an approach of using data mining techniques to assign employees to a suitable job. While their approach is based on current employees of an organization, the paper presents

Á. Rocha et al. (Eds.): WorldCIST'18 2018, AISC 745, pp. 194–202, 2018.
https://doi.org/10.1007/978-3-319-77703-0_19

interesting findings regarding the application of classification algorithms and their evaluation. Their focus is on application of decision trees and decision forests.

Azar et al. [6], use similar techniques based on decision trees but applied towards recruiting of new candidates. They define a 'promotion assessment' variable and evaluate what the main attributes that need to be considered to predict better this 'promotion score' target are. They can successfully reduce the dimensionality of their dataset into five important variables that need to be considered during the hiring process: province of employment, education level, exam score, interview score and work experience. A limiting factor of this research is the nature of the data they have obtained. More specifically, their data is consisted of a collection of records from Commerce Bank of Iran, from 2005 till 2006. As such the dataset is only limited to a specific domain and one can safely assume that all candidates share similar characteristics.

Chien and Chen in their research [1] aim to create a model relying on rule-based classifiers that would later be used to assist recruiters during the recruitment process. Their contribution is regarding define HR strategies that will be used during the recruitment process.

Our approach discussed later in this paper, shares similarities with the research published by Giri et al. [2], where they try to analyze data from web-based systems such as Twitter, LinkedIn and GitHub. Their approach proposes data mining approaches (mostly clustering techniques) that aim to allow recruiters are filtering mechanisms during candidate selection.

Finally, Laumer et al. [9], and Diaby et al. [10] proposes the implementation of job recommender systems using social media as well as integrated into recruitment systems. In [10] the authors show great improvements in recommending jobs through SVM.

Researchers have also analyzed the industry requirements for new jobs compared to the real supply in the market. One major issue is the over-qualification problem [11]. The over qualification problem is when a candidate is overqualified or too good for the job. In the labor market, companies are faced with individuals with same qualifications but are having different skill levels, so they can be overqualified regarding formal qualification, while the skills they possess are appropriate for the job requirements [12]. The skill mismatching has become a serious and growing concern for many companies and also for employees. The over qualification has significantly received more attention than the one of under-qualification (under-skilled), because of fears that it may have been caused by the increased supply of universities and graduates [11].

For this research, we have direct access to data generated by web-based commercial platforms that aim to connect recruiters with job seekers. The entire analytical platform of these systems was developed as part of this research, and this paper presents some of the findings of the same.

2 A Job Platform Data Warehouse

As previously mentioned, this research is based on data generated by two platforms, vYou [7] and CVManager [8]. Both are implemented for the Swiss job market, but deployments exist in other countries such as Germany, Austria and Macedonia. The first one is a system that facilitates the interviewing processes with job applicants.

Implemented as a web and mobile-based system, the platform allows companies to schedule 'offline' and 'online' interviews through the WebRTC protocol. Offline interviews are interviews for which the questions are pre-recorded, and the candidate can answer them at their convenience. Online interviews are in principle video conferencing rooms, where the candidate can attend a meeting with one or more recruiters and be interviewed classically. During the interview, all the interaction between parties is recorded. Such recordings include the entire video communication, time-stamped notes registered by recruiters as well as any textual communication between recruiters. The interviewing process can go through many phases until a candidate is selected; in every phase, each interview is attached to the candidate file and can be reviewed by recruiters at any time. Also, job seekers can use this platform to maintain their profile that can later be shared with companies.

The second platform, CVManager is a platform that allows companies to publish vacancies and handle all other recruiting processes that are not related to video interviews. This platform allows seamless integration with job posting sites, career sites of the companies as well as manages social media marketing. Furthermore, it allows communication with the vyou platform from where it loads candidate profiles and does automatic vacancy/candidate matching.

Companies that are recruiting new employees need to know whether their posted vacancies are performing satisfactorily. We have identified the following KPI as the most important factors:

- Number of applications per vacancy
- Demographic drill down of applicants by - gender, age, city, etc.
- Competency of applicants - skills, professions, education, languages spoken, etc.
- Acceptance rate analysis – number of applications vs. number of hires, number of invited applicants per vacancy
- Number of interviews
- Time to hire
- Time per review
- Time to interview/time for interviews
- Number of vacancies per company/number of vacancies posted on external platform
- Referral analysis – from where the candidate applied (directly through CV Manager, Google, other recruiting sites, the website of the company …)

Most of the above KPI are self-descriptive and answered through descriptive statistics. The number of applications is being tracked automatically by the system. We have implemented different granularity levels, such as applications by age, location, time of day, gender, etc. Based on such information, we can give suggestions to companies that will help them be more effective when publishing job adds.

The fourth KPI is related to the quality of applicants. It represents the ratio between the number of applications vs. the number of candidates invited for an interview. When supported by the number of candidates accepted for a vacancy, a company can get valuable information regarding both the quality of applications and the quality of the job vacancy. One can quickly identify when an open vacancy is poorly defined and propose improvements to the recruiter/company.

Time to hire represents a set of variables that evaluate the performance of a vacancy. It provides information to recruiters related to the time needed to hire a candidate (after the vacancy is published), the time before the first applicant has applied, time until the first interview is scheduled, etc. With such information, companies can at least estimate the amount of time needed until a vacancy will be filled. Similarly, time per review represents the time needed to review a candidate application.

To accommodate all the needs of companies, by answering all the above KPIs as well as to provide more advanced analytics, we have designed an analytics platform that acts as a data warehouse which records all information that is not needed in real-time and will be used for future purposes. This warehouse (see Fig. 1) is part of the analytics package offered by the platform. Data is stored in a relational database management system as well as in a document store for some larger scale datasets.

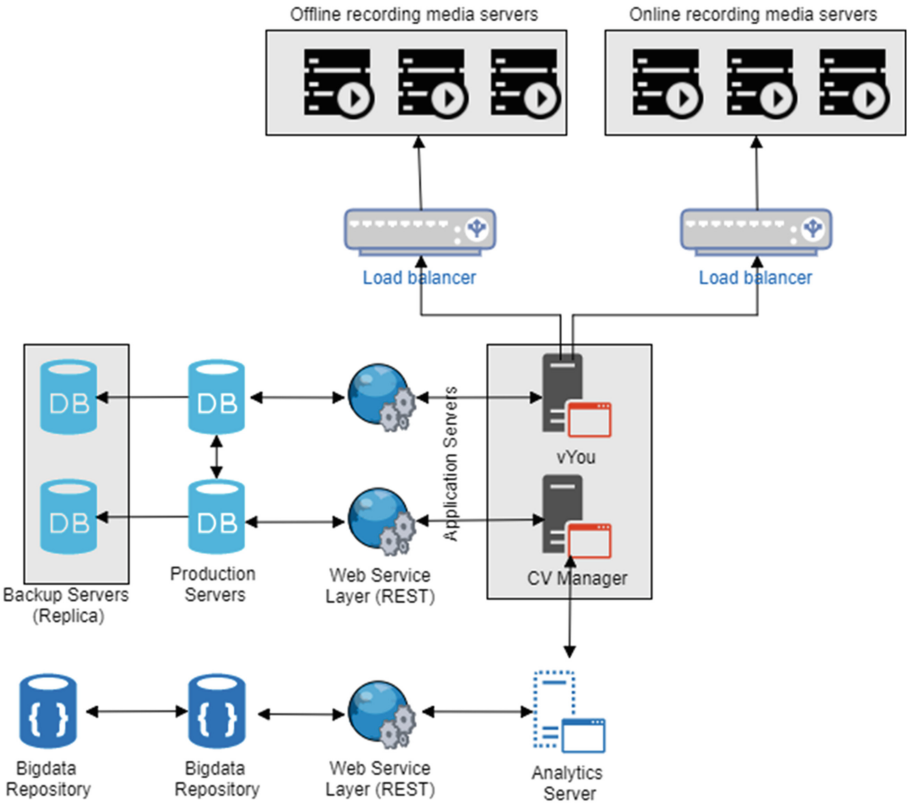

Fig. 1. Overall system architecture

3 Predicting the Job Market Needs

The needs of the job market are in a constant fluctuation. Recruitment companies need to understand the trends and be able to predict future movements of the market. For this purpose, we have established some mechanisms. These mechanisms are based on data mining algorithms, e.g., application of the method of least squares for anticipating skills required by the industry, or using clustering methods for building job recommendation systems.

To implement data mining algorithms, a data warehouse was implemented that would store historical records for all activities facilitated by the two systems we have implemented. The data warehouse is part of a larger analytics server and stores both relational and non-relational data. The non-relational data are of greater interest here, as they are represented in the form of events with specific timestamp. Such events range from login/logout information, publishing, viewing, applying for a vacancy, recording skills, and skill requirements, an update of user profiles, etc.

3.1 Skill Management

The analytics platform relies greatly on analysis done related to candidate skills as well as vacancy requirements (again represented in the form of skill requirements). For this purpose when a vacancy is posted, companies are required to register skills needed from candidates and rate them on a scale from 1 to 5. The same is for candidates, when they design their profiles, they can enhance them by registering and rating their skill-set. It is clear that the rating process can be subjective, and during clustering, we have also considered this.

Technically, every new vacancy is associated with a Skill vector, consisted of key-value pairs <skill, level>. A very similar data structure is associated with applicants. Since the information is time stamped, any new changes (especially when candidates update their profile), this new skill addition provides valuable information regarding new trends in the job market as well as candidate development.

One major problem we have practically faced was the case of diverse spelling of skills. This can happen both due to misspelling or different standards of spelling. The following example shows the scale of the problem. For IT Skills, we see in our database the following skills (from different companies): *EDV; EDV Kenntnisse; EDV Kentnisse; EDV-Kenntnisse; IT; IT Skills; IT Knowledge*. All of these entries represent the same concept, and we had to find means to alleviate this problem. To handle incorrect spelling, new versions of the system use UI tips to recommend skills to users from a list of more than 72,000 different skills. If a new skill is added, then it has to go through an approval process. Before being manually approved, the system will compare this new entry with all the skills in the database for similarity, using Levenstein distance. This allows correction of minor spelling errors and has proven to be very effective for us. The manual process of skill approval also allows administrators of the system to group diversely spelled skills together. This way all the above mentioned skill, belong to the same category: IT Skills.

3.2 Analyzing Trends in the Job Market

One of our research questions was if we could predict skills that would be required in the job market in the future. Our database is designed to track appearance and trends of skills in various periods of time. This way companies can predict seasonal workers, skills that might be required in the future as well as give weight to candidates.

By giving weight to candidates, the system actually rates candidates by their importance in the job market, as a function of the rarity of skills that a candidate possesses. We can easily calculate this by analyzing skills from vacancies and comparing them with individual skill presence in candidate profiles. When a candidate with a rare skill applies for a job, we inform the company that this candidate is a VIP candidate. This is also very valuable to recruiting companies as their business logic does not rely only on the quantity of candidates they hire but also on their quality.

In addition, we try to also help candidates develop news skills based on the needs of the job market or also remind them update their profile with a skill the system predicts the candidate already possesses. The system also gives suggestions to companies about skills in demand and helps them devise staff development programs.

A simple analytics would be to evaluate one skill requirement. The chart in Fig. 2 shows the percentage of time IT Skills are required in the accounting industry in Switzerland from January till November of 2017. Usually, this skill is often grouped together with other skills, related to vacancies posted. In the accounting industry, we often see it grouped with: *[excel, englisch, fakturierung, führungserfahrung ...]*. As time progresses, new skills might start appearing in this group of skills. This could be due to introduction of new work position or a seasonal job requirement.

This could be best illustrated with a deployment of our platform for the career center of South East European University in Macedonia. The IT industry in that country is predominantly Java or .Net based. Thus skill requirements usually are from those areas *[J2EE, C#, SQL Server, MySql, PostgreSql, jQuery, Angular ...]*. One noticeable spike of new skills requiring *NAV Developers* was noticed in October of 2016 and continued until June of 2017. It was determined that this was due to the establishment of a major company working with Microsoft Dynamics NAV in the country.

The system can quickly determine such new shifts by applying Naïve Bayesian classification technique. Note that upon vacancy creation companies classify the vacancies to their appropriate category, thus giving class information. Candidates however, do not have such categorization. Prior probability of a skill is essentially the frequency of that skill appearing as a requirement in vacancies. The probability of a candidate belonging to a class is the posterior probability of the product of all skill probabilities as seen in (1).

$$P_{(C|X)} = \prod_{i}^{n} P(s_i|C_i) \tag{1}$$

As new skills start appearing in the corpus of skills for that class, at first their relevancy can be simply discarded. However, when their frequency improves, thus increasing their probability, the skill will gain relevancy and affect how customers will be classified. This information is used to classify candidates and recommend vacancies to them, but is also used to early detect new skills and monitor the trends of their relevancy growth.

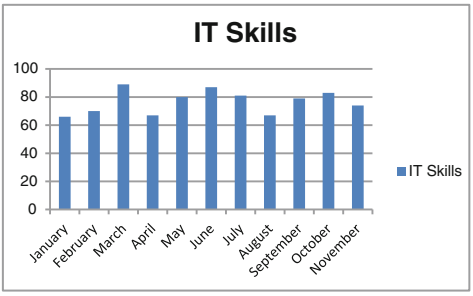

Fig. 2. IT Skill requirement in accounting vacancies

3.3 A Job Recommendation System

As discussed earlier, we can see that we already have skill vectors for both candidates and vacancies. Being that the data structures are equal, we can apply distance metrics to match candidates with corresponding vacancies. For privacy issues, we only provide job recommendations to candidates seeking for jobs; thus avoiding recommending suitable candidates to companies that have published a vacancy. To calculate the distance we evaluated three approaches. All three of them were subject to usability testing by clients in real-life scenarios, and the third approach was chosen as the better-accepted solution.

The first approach uses simple Euclidian distance; it calculates the distance between a vacancy V and candidate C through as in (2).

$$d_{V,C} = \sqrt{\sum_{i}^{n} (V_i - C_i)^2} \tag{2}$$

Results are kept in a candidate-centric distance matrix, which is updated whenever a new vacancy is added or whenever candidates update their profile by introducing a new skill or updating the rating of an old skill.

However, there are two main problems when using the above approach. The first problem is related to the subjective nature of the skill rating. Two candidates can have different rating criteria. The second problem is that in this approach jobs for which a candidate is both over-qualified and under-qualified are treated as equal.

To solve the subjectivity issue, we use the Jaccard similarity coefficient approach (3) and only treated skills as binary variables, ignoring rating. Therefore, a candidate

having skill level 5 in Java would be the same as another candidate having skill level 1 same in Java.

$$S_{V,C} = \frac{M_{11}}{M_{11} + M_{10} + M_{01}} \tag{3}$$

Finally, over-qualification is solved by slightly altering the Euclidian distance matrix as in (4). Using the cube root, we lose the symmetrical property of the Euclidian distance.

A usability study, conducted with eight companies in Switzerland (publishing more than five vacancies per week) and 50 job applicants concluded that the third approach was more acceptable.

$$d_{V,C} = \sqrt[3]{\sum_{i}^{n} (V_i - C_i)^3} \tag{4}$$

4 Conclusion and Further Discussion

In this paper, we presented our research that aims to improve the recruitment processes through data analytics. We present the major findings of our system, starting from simple statistical means to improve recruitments and vacancy posting. The paper continues discussing more advanced analytical approaches that aim to either predict job market fluctuations and to develop a recommender system. All these findings are successfully implemented and tested in a production environment.

The systems can and need to be further developed. For one we are experimenting with Deep Learning methods that aim to increase the machine learning capabilities of our systems. One application evaluated is the usage of Kohonen Self Organizing Maps, with which we will cluster applicants and improve the recommendation system. Another field of research we are investigating is the application of computer vision algorithms for the purpose of detecting emotion during the interviewing processes.

We also have started improving the recommendation system through integration of external data. Among them is commute time needed for employees based on candidate address and job schedule. This information is retrieved through Google Api.

Finally, being aware of the power of data mining and machine learning, we need also to consider implications in the privacy of our users. While we take great precautions to respect users privacy, we often find ourselves walking a tightrope. One of such cases is the recommendation system, where it can be clearly seen that the same algorithms used to recommend jobs to candidates can be used to recommend candidates to companies. However, this information can be used inappropriately, by spamming candidates or head-hunting employees of other companies.

References

1. Chien, C.-F., Chen, L.-F.: Data mining to improve personnel selection and enhance human capital: a case study in high-technology industry. Expert Syst. Appl. **34**(1), 280–290 (2008). Elsevier
2. Giri, A., Ravikumar, A., Mote, S., Bharadwaj, R.: Vritthi - a theoretical framework for IT recruitment based on machine learning techniques applied over Twitter, LinkedIn, SPOJ and GitHub profiles. In: 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, pp. 1–7 (2016)
3. Javed, F., Luo, Q.: Carotene: a job classification system for the online recruitment domain. In: 2015 IEEE First International Conference on Big Data Computing Service and Applications (2015)
4. Singh, S., Kumar, V.: Performance analysis of engineering students for recruitment using classification data mining techniques. Int. J. Sci. Eng. Comput. Technol. **3**(2), 31 (2013)
5. Jantan, H., Hamdan, A.R., Othman, Z.A.: Towards applying data mining techniques for talent mangement. In: International Conference on Computer Engineering and Applications, vol. 2 (2011)
6. Azar, A., Sebt, M.V., Ahmadi, P., Rajaeian, A.: A model for personnel selection with a data mining approach: a case study in a commercial bank. SA J. Hum. Resour. Manag. **11**(1), 10 (2013). Tydskrif vir Menslikehulpbronbestuur, Art. #449
7. Marseco Software: vYou platform. https://www.vyou.ch. Accessed Nov 2017
8. Marseco Software: cvmanager platform. https://www.cvmanager.ch. Accessed Nov 2017
9. Laumer, S., Eckhardt, A.: Help to find the needle in a haystack: integrating recommender systems in an it supported staff recruitment system. In: Proceedings of the Special Interest Group on Management Information System's 47th Annual Conference on Computer Personnel Research, pp. 7–12 (2009)
10. Diaby, M., et al.: Toward the next generation of recruitment tools: an online social network-based job recommender system. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 821–828 (2013)
11. Quintini, G.: Over-qualified or under-skilled: a review of existing literature. OECD Social, Employment, and Migration Working Papers, No. 121, OECD Publishing, Paris (2011)
12. Green, F., McIntosh, S.: Is there a genuine under-utilization of skills amongst the over-qualified? Appl. Econ. **39**(4), 427–439 (2007)