

Semi-automatic Tool for Parsing CVs and Identifying Candidates' Abilities and Competencies

Alexandra CERNIAN*, Dorin CARSTOIU and Bogdan MARTIN

University Politehnica of Bucharest, 313 Splaiul Independentei, Bucharest, Romania

* Corresponding author

Keywords: Data mining, Semi-automatic tool, Recruitment, Semantic technologies.

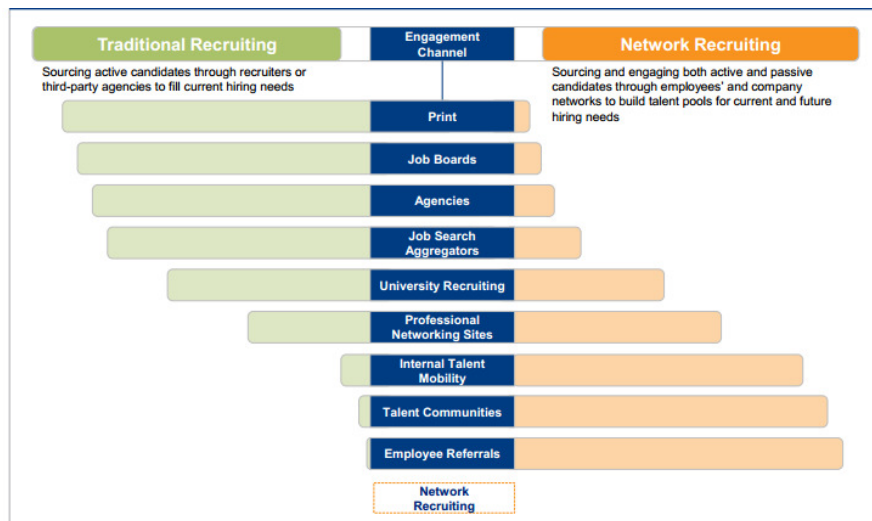
Abstract. This paper presents the design and implementation of a semi-automatic tool able to analyze a set of CVs in order to identify a suitable candidate for a particular job. The process is based on semantic processing of the resumes and matching their associated scoring with user-configured dictionaries. The purpose is to reduce the recruiter's processing time by eliminating certain repetitive activities in the resume analysis procedure and to obtain a qualitative improvement by highlighting the competencies and qualities of candidates based on complex and customized semantic criteria.

Introduction

The most important role that managers and leaders have in a business is to hire the right people for the right job. If we cannot figure out how to find and evaluate the appropriate candidates and manage to convince them to join our organization, progress will be difficult.

Currently, the recruitment process has become very dynamic and is in constant change. Besides finding and evaluating candidates, companies must differentiate between them through their own brand, the competencies, values and experience of the candidates and a good relationship management process with candidates.

The recruitment process can be seen from two different perspectives, one traditional and the so called "network recruiting" [1]. In terms of use, traditional recruitment is the most common and involves identifying people by focusing on printed materials or online recruitment agencies (see Figure 1).



Source: Bersin by Deloitte, 2014.

Figure 1. Network recruiting.

An improved recruitment process focuses on a better review process for resumes in order to efficiently identify the suitable candidates for a particular job. A purpose here would consist of reducing the recruiter's processing time with the use of tools that eliminate certain repetitive activities

in the resume analysis procedure and a qualitative improvement by highlighting the competencies and qualities of candidates based on complex and customized criteria [2]. One such example is the aggregation of data from a large number of such documents to extract organizational skills based on the candidates' professional experience.

It is important that such an application implement one or more of the following functionalities for automatic parsing of resumes regardless of their structure:

- Ability of parsing documents for a large number of document types (pdf, rtf, etc.)
- Ability to perform complex searches in the candidates' database
- Ability to generate a shortlist for a particular job description
- Ability to define and constantly update job profiles
- Ability to display reasoning algorithm used by the recruiter and permanently improving it

Semi-automatic tool for parsing resumes. Overview.

This paper presents the design and implementation of a semi-automatic tool able to analyze a set of CVs in order to identify a suitable candidate for a particular job. The process is based on semantic processing of the resumes and matching their associated scoring with user-configured dictionaries.

Figure 2 depicts the flow of the process.

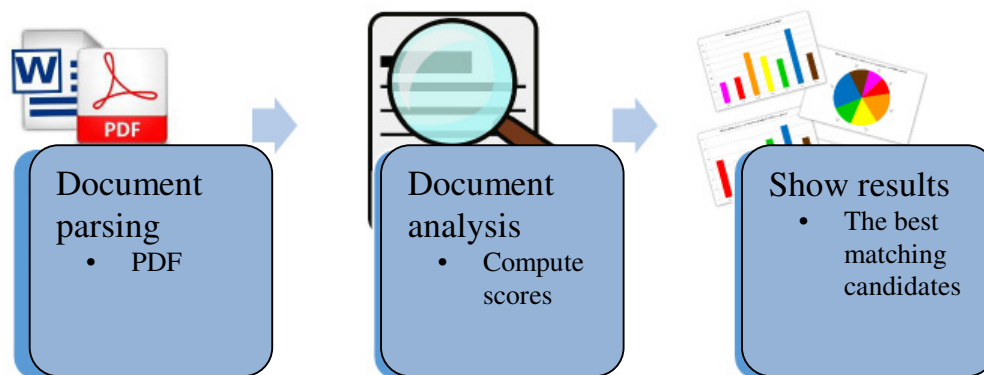


Figure 2. Application flow.

The main objectives of the application are the following:

1. Implement a semi-automatic search model used for parsing candidates CVs
 2. Implement a customizable module where HR specialists can configure the profile they look for among their candidates.
 3. Implement an algorithm to find the candidates that best match a specific job based on their CV.
- The candidates' selection algorithm starts each search process based on the configurations made by the user and takes into account the following criteria:
- The type of job and main competencies required
 - The candidates' education, based on a classification of educational institutions and the professional competencies they provide
 - The work experience of a candidate and a classification of their abilities and competencies

The CV Parsing Module

This component of the application processes information in four main phases:

1. Get input data from the user (one or more CVs) and transforming them into text files
2. Parsing the CV in order to split the information into individual sections: personal details, education, work experience etc.
3. Classify specific words in each section in order to leverage the candidates' selection model

4. Store the information in a database. The structure of the database is depicted in Figure 3.

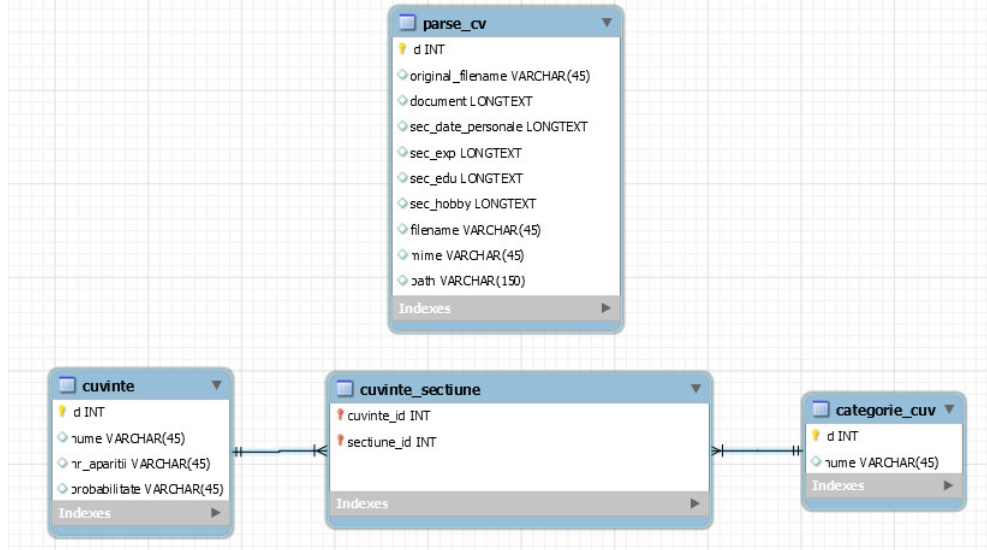


Figure 3. CV Parsing Database.

The text from the CVs is extracted with an open source library called Apache PDFBox [3], which facilitates document management, providing functions to create documents and extract content from existing documents.

The Candidates Analysis Module

This module has the following main components and functionalities:

1. *The identification component*: processes the CV and provides information to the CV Selection Module in order to determine the candidates that best suit a specific position.

2. *The abilities and competencies component*: (Figure 4) this component is based on a “abilities and competencies dictionary” (taxonomy), which is configurable by the user (recruitment specialist) and is constantly updated as CVs are analyzed.

3. *The companies and educational institutions classification component*: this component is based on two dedicated dictionaries (taxonomies) and has the purpose to provide a classification of the abilities and competencies that the candidates have acquired during their studies and work experience. The dictionary components are based on the SentiWordNet semantic taxonomy [4].

Janus

Identificare

Parse CV

CV-uri

Competente

Experienta prof

Experienta educ

Date Agregate









Cuvinte

Bogdan Martin

Dictionar Competente

Cauta

Tasteaza aici ...

ID	Tip	Grad Incredere	Termeni	Descriere	Map	UnMap	Edit	Delete
1	competenta	2.2	excel#1,advanced#2,pivot#3,adaptable#4	Can use excel at an advanced level communication adaptable; (Can use all the features of the excel for day to day usage.); (Can use table design.)				
2	abilitate	1.6	communication#1,proficient#2,pcm#3	Has proficient communication skills; (Has knowledge of PCM)				

+ Adauga

Figure 4. Competencies dictionary.

4. *The semi-automatic learning component*: this component uses the user input and the application generated classification to learn and generate new information, in order to update the dictionaries used by this module.

Figure 5 shows the database tables associated with the dictionary (taxonomy) components:

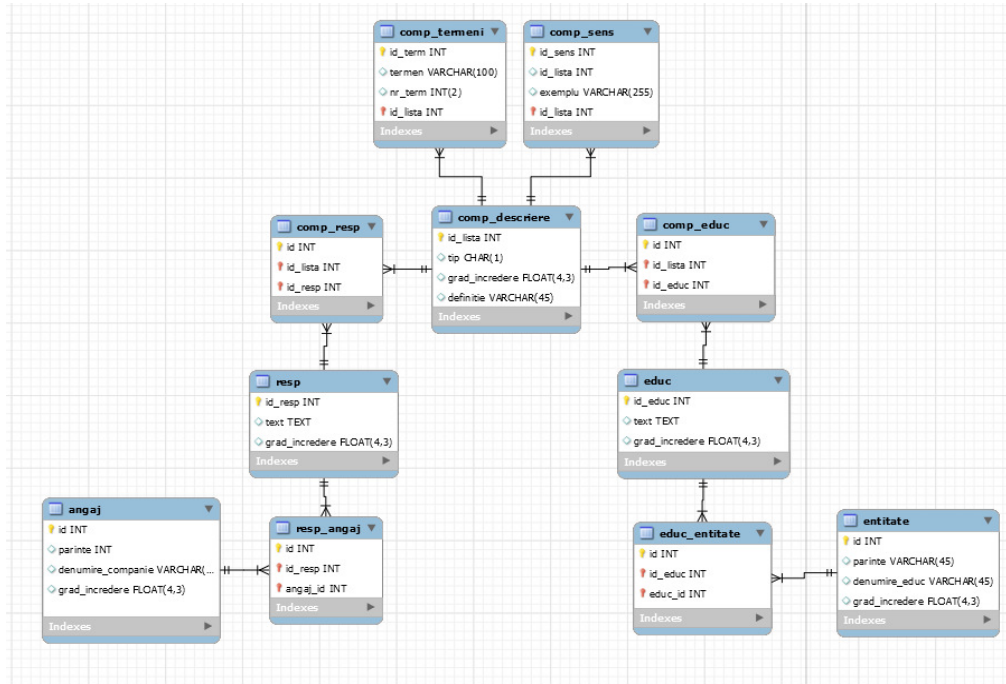


Figure 5. Dictionaries database tables.

At this point, the CVs are parsed and specific key terms have been identified and classified. The next step is to compute a matching score for each CV, in relation to the requirements for a specific position in a company.

The CV Selection Module

The score associated with a CV and its position between all the analyzed CVs is based on the number of identified terms that correspond to the search criteria, such as skills, job title, keywords or job responsibilities, as well as educational items related to employers and educational institutions in relation to the required skills.

The CV selection phase has several steps, as follows:

1. Initialization of the library used for text parsing.

The first step is initializing the library used to parse the texts in all sections of the CV and the definitions for competence, professional responsibility or educational background.

2. Take input from the user

The input consists of: competencies, abilities, free text, job title, professional responsibilities, education and interests.

3. Extract the data from the dictionary tables from the database.

4. Extract the previously parsed data from the database

- 4.1 For each CV

- 4.1.1. Compute score for each section

A section score is calculated by the number of appearances of elements that define the competencies or other required parameters

$$score_{section} = score_{section} + n \quad (1)$$

- 4.2. Compute score for CV

The score for the CV is calculated with the following formula:

$$score_{cv} = \sum score_{section} \quad (2)$$

5. Sort the CV by score.

6. Show the list of candidates based on the highest scores.

Conclusion

The application has been tested using a set of 150 CVs and 8 job descriptions and the results were 100% positive. The scores were accurately computed and the recommendations fitted the required profile for specific jobs. The semi-automatic tool for parsing resumes and identifying candidates' abilities and competencies leverages the selection process for recruitment specialists and reduces the time needed to manually process the information in the candidates CVs. Moreover, due to the integrated semi-automatic learning module, the competencies dictionaries are constantly updated and the selection model is improved. This tool can be a relevant decision support instrument for HR and recruitment specialists in identifying the best suited candidates for specific jobs, based on their profile and experience.

References

- [1] Josh Bersin, Predictions for 2015 Redesigning the Organization for a Rapidly Changing World, Deloitte, 2015: [http://www.cedma-europe.org/newsletter%20articles/misc/Predictions%20for%202015%20-%20Redesigning%20the%20Organization%20for%20a%20Rapidly%20Changing%20World%20\(Jan%2015\).pdf](http://www.cedma-europe.org/newsletter%20articles/misc/Predictions%20for%202015%20-%20Redesigning%20the%20Organization%20for%20a%20Rapidly%20Changing%20World%20(Jan%2015).pdf)
- [2] Hsiu-Fang Hsieh, Sarah E. Shannon, Three Approaches to Qualitative Content Analysis, Qualitative Health Research, Vol. 15 No. 9, November 2005 1277-1288, DOI: 10.1177/1049732305276687
- [3] PDFBox: <https://pdfbox.apache.org/index.html>, last accessed July 2016
- [4] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani, SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining: <http://sentiwordnet.isti.cnr.it/>, last accessed July 2016