Professor Paul POCATILU, PhD
E-mail: ppaul@ase.ro
Department of Economic Informatics and Cybernetics
Bucharest University of Economic Studies
Mihaela-Irina ENĂCHESCU, PhD Candidate
E-mail: irina.enachescu@csie.ase.ro
Bucharest University of Economic Studies
Alexandru DIȚĂ, PhD Candidate
E-mail: alexandru.dita@csie.ase.ro
Bucharest University of Economic Studies

ASSESSING A CANDIDATE'S SENIORITY LEVEL IN COMPUTER SCIENCE FIELD BY INTEGRATING SEMANTIC WEB TECHNOLOGIES WITH AUGMENTED REALITY

Abstract. *Usually, the process of recruiting employees is very complex, and this requires the allocation of numerous resources within the organization. Several steps of this process can be automated, especially related to Curriculum Vitae (CV) analysis. This paper presents the prototype of an application that combines semantic technologies with augmented reality (AR) in order to enhance the plain text from a CV content with context-aware information in regards to the technical background of the applicant. This approach proves to be more trustworthy than manual selection performed by the HR personnel, since rather than looking on an applicant profile as a collection of skills we try to reveal how all these competences interact with each other and what implicit knowledge the candidate has. We tested our prototype on a set of real CVs, gathered several metrics - number of found skills, number of knowledge graphs, number of significant knowledge graphs, maximum depth of known skills and primary known skills weight - and used them to analyze what influence they have on assessing the candidate's seniority level. We applied logistic regression with Bayesian approach and used the ROC curve in order to estimate the performance of the forecasting model.*

Keywords: *e-recruitment, semantic web technologies, graphs, ontology, regression, Bayesian approach, augmented reality.*

JEL Classification: L86, C88, Y80

1. Introduction

Recruiters have to screen dozens of CVs daily. While they can notice easily some key aspects such as previous experience expressed in years, stability in a

DOI: 10.24818/18423264/54.3.20.14

workplace, the ability to expose in a concise manner what the candidate knows to do best and even how organized a candidate is, when it comes to the actual knowledge the applicant possesses in the IT field, things can get more complicated. This is caused mainly by two reasons:

1) first, the recruiter has to have some background knowledge about the skills in the computer science domain and how they are related one with another,
2) it can be the case that although a candidate has not mentioned in the CV particular skills required by the job description, he or she can still be considered eligible, based on a more in-depth look at his background which can identify that there is a strong relation between what he knows and what the recruiter is looking for.

Obviously, we cannot pretend from a recruiter to know most of the skills in computer science field and nor to ask him/her to how they are linked with each other. However, we can agree that knowing all this information while screening a CV can bring advantages for both the company, as it will not lose a potential candidate while the candidate would have a fair chance to a job interview.

In order to solve this problem and to speed the recruitment process, while adding meaning to the candidate's knowledge encompassed in a CV, the authors of this paper propose a solution that will integrate semantic web with augmented reality technologies. The purpose is to augment the plain text CV content with context-aware information in regards to the technical background of the potential candidate.

The main objective of Augmented Reality (AR) systems is to provide additional information of a view with the aim to improve the user's perception about the reality. These applications use context-aware data (geographical position, recognized text, date-time, etc.) and user provided data in order to generate additional information. All this data and information needs to be linked with a formal semantics, which can be used to draw conclusions. That is exactly what semantic technology aims to address by labelling data with information related to its meaning and context. The result will be embedded in components for the AR system.

This paper continues a previous research presented in (Pocatilu et al., 2017) and (Enachescu, 2019).

The paper is structured as follows. Section 2 provides an overview about semantic web technologies, introducing the reader to its main concepts. Section 3 describes the concepts and the applications of Augmented Reality on mobile devices. Section 4 proposes a model that can assess the candidate's seniority level, based on the graph of skills, using a series of metrics collected from a set of CVs. In section 5 we introduce the architecture of a prototype system used to reveal to the recruiter the knowledge a candidate has in the IT field and how is that knowledge related to other significant skills in the same area. In section 6 we validate the resulting model and we estimate its forecasting performance. The

DOI: 10.24818/18423264/54.3.20.14

paper ends with the conclusions of our research and presents some future work steps.

2. Semantic Web Technologies

The expression "semantic web" was first proposed by Sir Tim Berners-Lee (Hitzler et al., 2010), an English computer scientist, known as the inventor of the World Wide Web (WWW). From the beginning, semantic web was coined as an extension of WWW, to allow computers to combine and intelligently process the content available on web, based on the meaning it has for people. To achieve this goal, storing the data using a machine-readable syntax, like HTML, is not enough. Semantic Web is providing "a basis for coding, exchanging, and reusing structured metadata among applications exchanging machine understandable information on the Web" (Ermilov et al., 2014).

Since 2001, several standards for semantic web formats, referred by the current terminology as Semantic Web Layer Cake (Ghosh et al., 2015), were published, trying to facilitate the exchange of rich sematic information, such as: Resource Description Framework (RDF), Web Ontology Language (OWL) – with OWL 2 enhancement since 2009, SPARQL Protocol and RDF Query Language.

Resource Description Framework, as presented in (Hitzler et al., 2010) and (Sicilia, 2014), is a formal language used to describe structured information. RDF is considered to be the main representation format for semantic web development. In opposition to HTML and XML formats, RDF does not only aim to correctly display documents, but rather to allow subsequent processing and recombination of their content. RDF is based on graph-oriented data schema. RDF graphs consist in a set of statements, each statement being composed of a RDF triple - "subject-predicate-object". The subject identifies the resource that the statement refers to, the predicate points to the subject's characteristic, as specified by the statement, and object represents the value of the property referred by the predicate. RDF is also employed to exchange metadata in specific application areas.

The world of ontologies is defined as one of abstraction and empiricism. As stated in (Sicilia, 2014, p. 95) "An ontology is an explicit, formal specification of a shared conceptualization". Also, in the same book, ontologies in the field of Computer Science are presented as "a model used to describe the world that consists of establishing a set of topics, properties, and relationships. A resemblance is established between the real world and the model created by ontologies".

OWL is a knowledge representation language used for creating web ontologies, which since 2004 become a W3C recommended standard. In 2009, OWL 2 was standardized. The main goal in OWL design was to find a reasonable balance between language expressivity on the one hand and scalability on the other hand. To give the user the possibility to choose between different levels of expressivity and reasoning capabilities, three OWL language variants were designed: OWL Full, OWL DL and OWL Lite (a straightforward comparison between these three dialects can be seen in (Hitzler et al., 2010). Advanced

DOI: 10.24818/18423264/54.3.20.14

ontology languages, such as OWL, support components like classes, properties, restrictions, relations, axioms, and individuals.

We have discussed about a few possibilities to represent information in a machine-readable way, like RDF that enables structuring and linking the information, and OWL that provides more powerful means for describing logical relations with significant complexity. Next step will be to effectively access this kind of information. While for relational databases SQL language is used for querying the data, in case of graph databases and RDF resources, SPARQL Protocol and RDF Query Language (in one word, SPARQL, pronounced "sparkle") is employed to retrieve and manipulate the data.

According to Sikos (2015), SPARQL is capable of:

- querying RDF files (local and online), Linked Open Data (LOD) datasets, and graph databases;
- building new RDF graphs, based on the queried graphs;
- adding or deleting RDF statements from a graph;
- inferring logical consequences;
- merging queries across distinct repositories;
- querying multiple data sources at once and dynamically merge graphs into a larger one.

The use of semantic web technologies proved notable results in many areas, such as: e-Learning (Bajenaru and Smeureanu, 2015), e-Recruitment, health, agriculture, bioinformatics, tourism and so on.

3. Augmented Reality in Mobile Applications

According to several authors, the AR presents two different approaches. Firstly, Weng et al. (2013) and Han and Zhao (2015) highlight that an AR system has the following characteristics:

- combines virtual and real elements - the generated elements are augmented on the device screen in order to make the user to feel that the virtual objects coexist with the real elements;
- it is registered in tree-dimensional space (3D) - the virtual information is displayed and aligned with the real-world objects;
- interactive in real time - the virtual objects created by AR systems presents some events which can be triggered by the users in the real time.

Dita (2016) highlights that Augmented Reality is part of Virtual Continuum, a context composed of Real Environment and Virtual Environment. The first element of Virtual Continuum contains all representations of real world. The second environment combines different components from virtual life. Moreover, Dita (2016) considers that ARs are closer to reality than virtual environment because this kind of systems extends the real objects with virtual elements. Likewise, Jamali et al. (2015) suggests that AR technologies extends the real world scene

DOI: 10.24818/18423264/54.3.20.14

with 2D or 3D computer generated objects and offers to the users the possibility to interact with them.

The existing literature (Geroimenko, 2012) suggests that augmented reality is divided in Location-Based and Vision-Based. According to Geroimenko (2012) and Karaman et al. (2016) the Location-Based consists in using of different instruments, such as GPS, gyroscope, compass or accelerometer, to estimate the location and rotation of the device in order to show the relevant information to the users. Secondly, Geroimenko (2012) argues that the Vision-Based consists in tracking different objects in the real life in order to display additional information on the device screen. Likewise, Luo (2011) considers that the Vision-Based processing of the AR systems, presents the following tasks: recognition and tracking. According to the authors, the scope of the recognition phase is to identify one or more pre-defined virtual objects in the video frame. Also, Luo (2011) suggests that the tracking task consists in estimation of the camera pose and augmenting the recognized virtual objects in video sequence continuously.

Other studies, such as Majid et al. (2015) consider that in the Augmented Reality technology a view of a reality can be modified by adding digital information on it in order to improve the person’s perception of the reality. In addition, the authors suggest that the AR systems have the following main components:

- Camera, which is used to collect the target information from the real scene.
- Marker, which encapsulates the target information that is necessary to AR system in order to augment the virtual elements.
- Mobile devices, which are responsible for storing and processing the content extracted from the image captured by the camera that contains the target information (marker).
- Digital content is the information which is displayed on the mobile screen when the camera is able to recognize and track the marker.

More studies, such as Sanchez et al. (2013) and Chen et al. (2016) present the benefits of using Mobile Augmented Reality (MAR) systems based on tracking elements. Sanchez et al. (2013) present a mobile augmented reality system based on visual recognition. According to the authors this process consists into tracking objects identified in real life with the camera of the mobile device and inserting the virtual elements, created by MAR technology, into the real world. Similarly, Chen et al. (2016) proposes a MAR system based on 3D registration technology. The same authors consider that the 3D registration consists in using the GPS and sensors from mobile devices in order to track the position and the pose of the devices in real time and real world and finally use this information to augment the virtual scene in the real life.

DOI: 10.24818/18423264/54.3.20.14

4. A Model to Evaluate the Candidate's Technical Background

We propose an ontology that expresses the relationships between skills within the computer science field. We will use this ontology to identify the relation between the other skills of a candidate based on his or her CV. Figure 1 presents a preview with a part of the designed ontology, comprising some of the classes and their instances.

The parent class of the ontology is Knowledge that has the following subclasses: Software development, Systems analysis, Hardware, Management etc., each of these having their own instances and/or sub-classes. For example, Scrum and Agile coaching are instances of class Management.

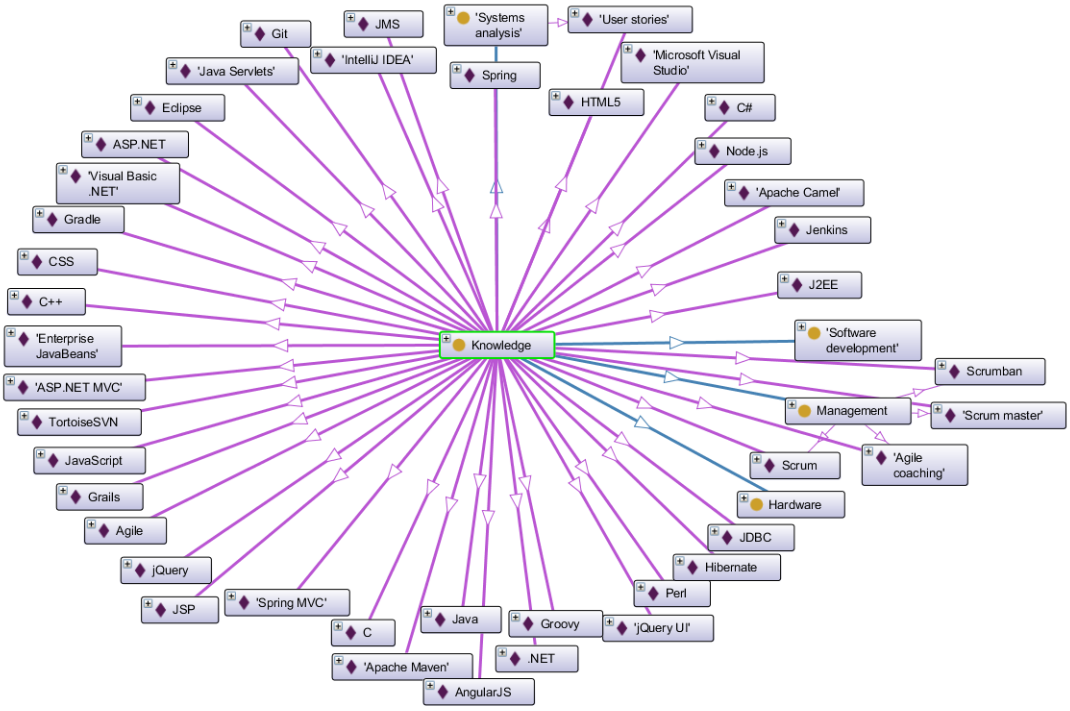


Figure 1. A preview with a part of the classes and instances from the proposed ontology (displayed in Protégé - OntoGraph)

Depending on the candidate's technical background, the skill-set can be comprised of one or more graphs. The more knowledge a candidate possess in a given area, the more primary nodes a graph contains. On the other side, for a candidate with basic knowledge in multiple fields, the illustration will be composed from multiple graphs, each having just one primary node.

The graph includes two types of nodes: main nodes and secondary nodes. Main nodes represent skills directly extracted from the candidate's CV. Secondary nodes consist in skills not mentioned in the CV, but extracted from the ontology, based on their relation with the main nodes.

Starting from the knowledge graphs that provide an overview of the candidate's competencies, we wanted to analyze if we can use them in order to assess the seniority level of the applicant.

Apart from the metrics collected from the candidates' knowledge graphs, we also evaluated the seniority using the years of experience (as illustrated in Table 1), based on the most common qualification levels, used in the IT industry, also presented in (AltexSoft, 2018). We appended also the seniority level as a last column in Table 2, for easiness in data centralization.

Table 1. Seniority levels based on years of experience in the field

Seniority Level	Years of experience
Trainee	less than a year of experience
Junior	1 - 3 years of experience
Middle	3 - 6 years of experience
Senior	over 6 years of experience

Our goal is to evaluate if any of those metrics have an influence on the seniority level and if they do then assert to what extent. The dependent/endogenous variable of the model is the seniority level. We considered also five independent/exogenous/explanatory variables:

- *Number of found skills* – represent the number of skills automatically extracted from the candidate's CV.
- *Number of knowledge graphs* – represent the number of connected components/sub-graphs in which any two skills (vertices) are connected to each other by paths. Having a path between two nodes indicate that those two skills are related. This information is extracted from the built ontology.
- *Number of significant knowledge graphs* – represent the number of knowledge graphs (as described previously) that are composed of at least three skills.
- *Maximum depth of known skills* – represents the length of the longest path, which crosses only main nodes (skills found in the CV).
- *Primary skills weight (%)* – represents the ratio between the number of main nodes in graph and the total number of nodes in graph.

Primary skills weight (%) is computed as follows: the knowledge graphs are sorted in descending order using the total number of main nodes in a graph. For graphs with the same number of main nodes, priority will have the ones with minimum total number of nodes. We note as $graph_1$ first of previously ordered graphs that should also have $\sum main\ nodes$> 1. Then:

DOI: 10.24818/18423264/54.3.20.14

$$Primary\ skills\ weight\ (\%) = \frac{total\ number\ of\ main\ nodes\ in\ graph_1}{total\ number\ of\ nodes\ in\ graph_1} x100$$

As stated previously, for the given data set, the dependent variable – seniority level – can record two values: middle and senior, which will be encoded with 0 (for middle) and 1 (for senior). Because the fact we aim to explain is measured by a qualitative categorical dichotomous variable, we can apply either a logistic regression or a probit regression. Although in most scenarios both models match the data similarly, since the function used to calculate the probabilities is distinct, their results are sometimes slightly different. As stated by Hahn and Soyer (2005) "*the conventional wisdom is that in most cases the choice is largely a matter of taste.*" Logistic regression is considered to be more popular because coefficients can be easily interpreted in terms of odds ratios, so it provides an intuitive manner to explain the effects (Grace-Martin K., 2018). This is the reason we decided to apply logistic regression for our analysis.

Logistic regression is part of Generalized Linear Model (GLM) class. Its goal is to find the best model that describes best the relationship between the characteristic of interest, in our case the seniority level, and a set of independent predictor variables (number of found skills, number of knowledge graphs, number of significant knowledge graphs, maximum depth of known skills and primary known skills weight). According to Gelder (2012), logistic regression is used to predict the probability that an event will occur.

The logistic regression equation is written as (Schoonjans, 2019):

$$\mathrm{logit(p)} = \mathrm{b_0} + \mathrm{b_1x_1} + \mathrm{b_2x_2} + \cdots + \mathrm{b_kx_k},$$

where p is the probability that a characteristic is present, $\mathrm{b}_{\mathrm{i}=\overline{1,k}}$ – unknown parameters that need to be estimated, $\mathrm{x}_{\mathrm{i}=\overline{1,k}}$ – the independent variables and k is the number of independent variables.

The odds are expressed as:

$$\mathrm{odds} = \frac{\mathrm{p}}{1-\mathrm{p}}.$$

In our case, p is the probability of the candidate to be senior, k = 5 and

$$odds = \frac{\text{probability of the applicant to be senior}}{\text{probability of the applicant to be middle}}.$$

The logit transformation is defined as the logged odds (Gelder, 2012), (Schoonjans, 2019):

$$\mathrm{logit(p)} = \ln(\frac{p}{1-p})$$

DOI: 10.24818/18423264/54.3.20.14

In order to perform the regression analysis we have used the R language. Among the software available for statistical data analysis, R is a powerful, open source tool that offers a wide range of regression analysis options, with dedicated packages for that purpose.

5. The architecture of the proposed solution

In order to achieve the proposed objectives, we developed an architecture that includes the following modules: content extractor, skills detector, knowledge builder and skillset illustrator.

The input of the application is the CV of a potential candidate. This can be:

- a directly uploaded CV, when we use the desktop application;
- scanned and viewed using the camera of a mobile device.

Figure 2 depicts the architecture of the proposed solution.

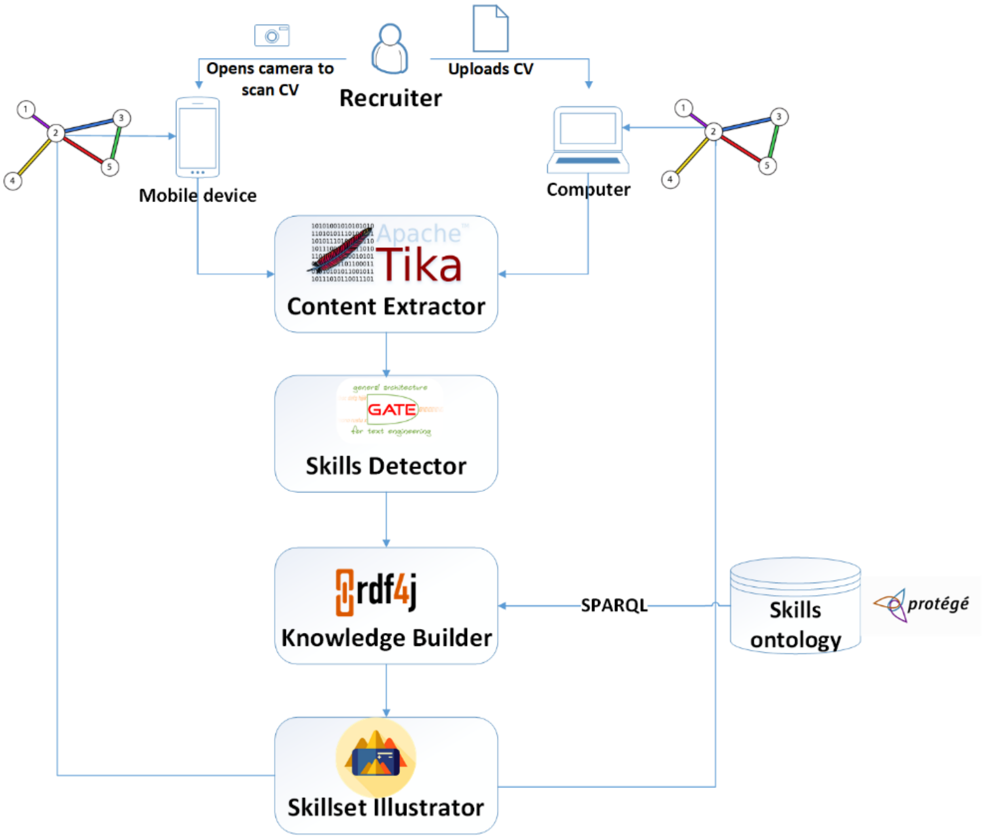


Figure 2. The architecture of the proposed solution

DOI: 10.24818/18423264/54.3.20.14

The **content extractor** is using Optical Character Recognition (OCR) or content extraction techniques to parse the text from the CV. OCR is applied when we are using the mobile device and we capture in frames the content of the CV. In order to extract the text from the picture, Tess-Two library is used. This is an open source library that combines capabilities from both Tesseract and Leptonica libraries, to recognize text written in multiple languages from an image (Bloomberg, 2014). In case the CV is uploaded as a .pdf or .doc/.docx file, we use Apache Tika toolkit. Apache Tika consists in a set of tools designed for extracting content and metadata from different types of documents, like HTML, XML, documents generated using Microsoft Office suite (Word, Excel and PowerPoint), PDF, RTF or even multimedia files, like JPEG, MP4 etc. All these documents types are being processed through a single common interface, called Parser, making Tika a powerful and versatile text analysis tool (Thai, 2019).

In the next phase, the text will be sent to a **skills detector service** that will be responsible to parse the content and identify the skills the candidate possesses. In order to achieve this, skills detector service is using an open-source infrastructure, called GATE (General Architecture for Text Engineering) that provides tools for developing and delivering software components that process human language. GATE is distributed together with an information extraction system, called ANNIE (a Nearly-New Information Extraction System), that consists in a predefined collection of algorithms, used for creating RDF or OWL representation from an unstructured content, based on semantic annotations. In order to extend the annotations created using ANNIE, JAPE (Java Annotation Patterns Engine) is also employed. JAPE is a pattern matching language developed specially for GATE that creates rules for adding new annotations using the concept of grammar (Cunningham et al.,2017).

After obtaining the list with the technical competences from the CV, we will query for each one the designed *ontology* in order to identify its relation with other skills. RDF4J, an open source Java framework, is used for parsing and querying the ontology. After this step, we get the list with the skills already present in the CV, but we will extended it also with other competencies that are related to the first ones. This is done using the **knowledge builder**.

In the end, the recruiter will see on the device screen, over the CV content, one or more graphs depicting the skill-set of the candidate. This is managed by the **skillset illustrator module**. The competences directly extracted from the CV will be drawn as main nodes and others as secondary ones. Edges in the graph will describe relations between skills.

After uploading the candidate CV for a given position into the application, the recruiter will be able to see a graph with the skills of the candidate and how those skills relate to others in the field. An example of the skill-set graph (as displayed in the web version) is presented in Figure 3.

In Figure 3 we recognize “Apache Camel” as primary node (skill found in the candidate CV, depicted in green and having a bigger size in the graph) and “JMS”

DOI: 10.24818/18423264/54.3.20.14

as secondary (depicted in blue, smaller size, not present in the CV, but connected to Apache Camel).

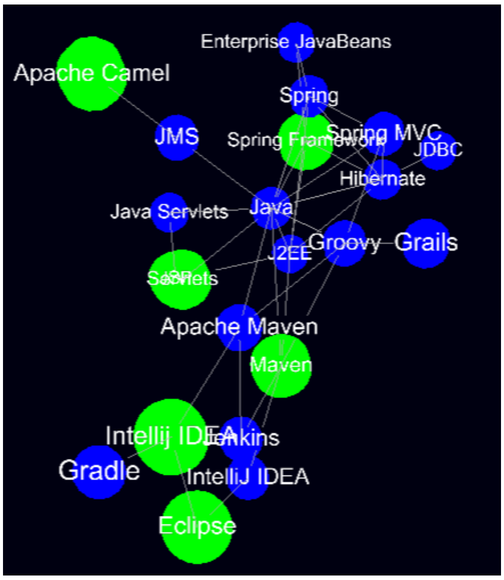
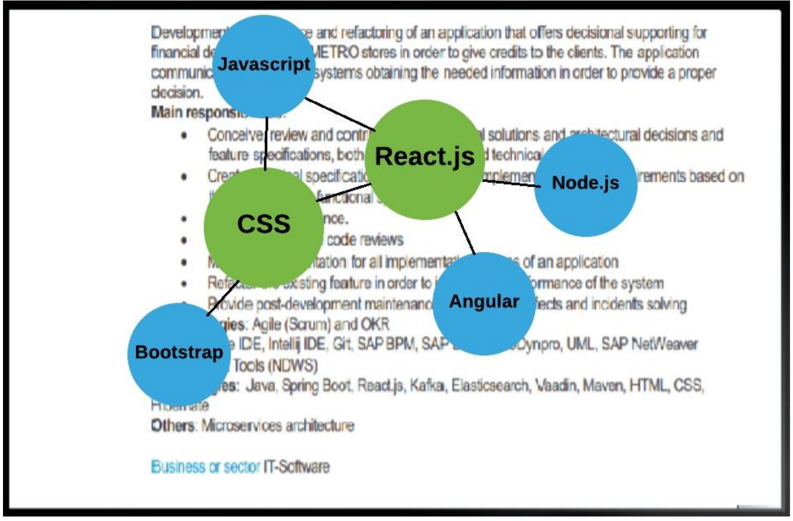


Figure 3. Graph containing the skills of a candidate

An example with the projection of the skill-set enclosed in a candidate CV, using AR, can be seen in Figure 4.



Developmen e and refactoring of an application that offers decisional supporting for financial de METRO stores in order to give credits to the clients. The application communic systems obtaining the needed information in order to provide a proper decision.

Main responsi

- Conceive review and contr l solutions and architectural decisions and feature specifications, both d technical
- Cre al specificatio mplemen rements based on functional
- nce.
- code reviews
- M tation for all implementa s of an application
- Ref xisting feature in order to ormance of the system
- Provide post-development maintenanc fects and incidents solving

gies: Agile (Scrum) and OKR

e IDE, Intellij IDE, Git, SAP BPM, SAP Dynpro, UML, SAP NetWeaver Tools (NDWS)

es: Java, Spring Boot, React.js, Kafka, Elasticsearch, Vaadin, Maven, HTML, CSS, Hibernate

Others: Microservices architecture

Business or sector IT-Software

Figure 4. Skill-set graph projected over CV using AR

One thing worth to be mentioned here is the difference between using the mobile versus desktop form of the application. When we rely on the mobile phone

DOI: 10.24818/18423264/54.3.20.14

camera to obtain the CV content, we will get parts of the CV, as the camera will be able to get frames with one page of the CV at a time, but not all pages simultaneously. In order to solve this we need to ask the user to confirm if next collected frames are part of the same CV and, in this case, instead of creating a new graph with the skillset, we will just enrich the current one. In this manner, we ensure that we deliver to the recruiter a picture with the full technical knowledge of the candidate, comprised in his CV. On the other hand, when we upload the CV into the application, we will analyze in one shot all its content, so we avoid providing the recruiter with an incomplete skillset.

6. Results and discussions

In pursuance of this analysis, we collected the CVs from the applicants for a Senior Java Developer position fora company from Bucharest. Taking into account the average number of applicants for a similar position in the company, we gathered 20 CVs and uploaded them into our application. For each candidate we collected a series of metrics we considered relevant for explaining the seniority level that can be found in Table 2. All the applicants are either middle or senior, as expected for a Senior Java Developer position.

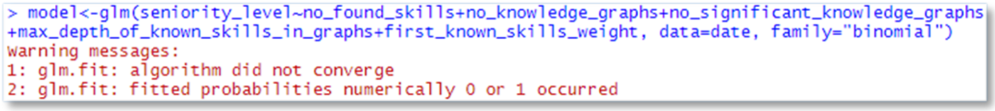
Table 2. Metrics collected from the candidates' knowledge graphs

Candidate	No. of found skills	No. of knowledge graphs	No. of significant knowledge graphs	Maximum depth of known skills	Primary skills weight (%)	Seniority level
C1	13	7	3	4	27.272727	MIDDLE
C2	15	8	4	5	31.25	MIDDLE
C3	32	19	3	7	34.482759	SENIOR
C4	25	13	4	6	33.333333	SENIOR
C5	46	28	5	8	34.042553	SENIOR
C6	37	15	2	7	45.16129	SENIOR
C7	21	11	2	7	29.166667	SENIOR
C8	14	6	3	4	29.166667	MIDDLE
C9	17	10	2	7	30.769231	SENIOR
C10	14	5	2	3	30.769231	MIDDLE
C11	17	7	2	6	27.586207	MIDDLE
C12	14	7	4	6	37.5	SENIOR
C13	17	8	2	6	32.142857	SENIOR
C14	11	9	3	2	16.666667	MIDDLE
C15	15	8	2	5	29.166667	MIDDLE
C16	12	6	3	5	26.923077	MIDDLE
C17	14	8	1	6	25.925926	MIDDLE
C18	16	11	3	4	30.769231	MIDDLE
C19	11	8	2	3	21.052632	MIDDLE

DOI: 10.24818/18423264/54.3.20.14

Candidate	No. of found skills	No. of knowledge graphs	No. of significant knowledge graphs	Maximum depth of known skills	Primary skills weight (%)	Seniority level
C20	20	12	1	5	30	SENIOR

We used the proposed model and we applied *glm()* function with binomial family, thus indicating the response variable distribution function, on the analyzed data set.



```
> model<-glm(seniority_level~no_found_skills+no_knowledge_graphs+no_significant_knowledge_graphs
+max_depth_of_known_skills_in_graphs+first_known_skills_weight, data=date, family="binomial")
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

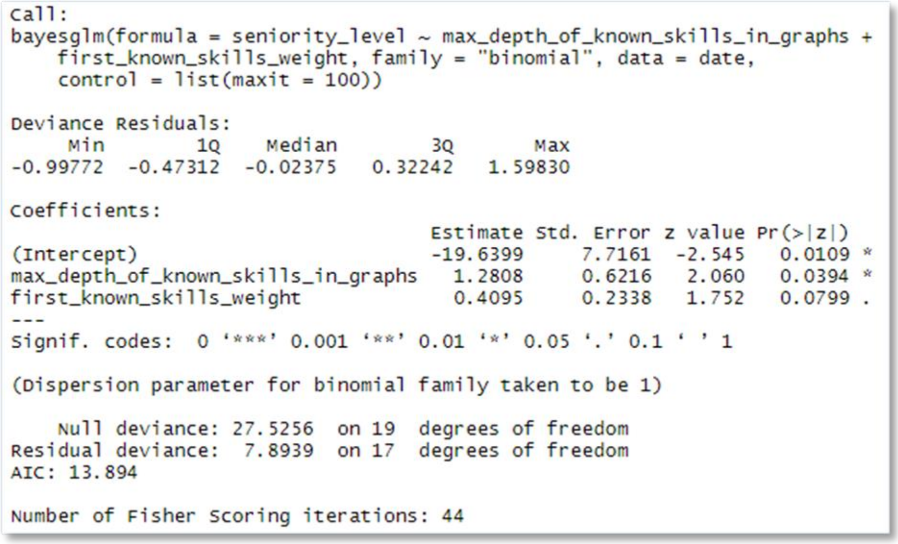
We notice the generated warning message that the probabilities predicted are 0 or 1. We obtained high values for the estimated coefficients of the model for several independent variables, with the associated standard errors increased and p-value (probability value to quantify the statistical significance of the parameters) for all coefficients of 1. All of these point out that no parameter is statistically significant and thus the model is not valid for interpretation. All previous observations indicate that we are facing a problem of quasi-complete data separation, a situation that is quite common in the case of logistic regression. Complete/quasi-complete data separation occurs when the response variable separates an independent variable or a combination of the predictor variables completely or to a certain extent, in the case of quasi-complete separation. The volume of the observations set (20 CVs) is a contributing factor to this problem.

In the case of a quasi-complete separation problem, the simplest solution strategy is to exclude the independent variables that have led to this situation. In the present case, this is not a correct approach because we have several independent variables that have generated this problem and which are important in estimating the seniority level. A method commonly used to correct the separation problem and the reduced sample size of logistic regression data is the Bayesian approach. In R, this can be achieved using *bayesglm()* function.

In the current case, having 5 predictor variables, there are 2^5 models we can choose from. We applied Akaike Information Criterion (AIC), which assesses both the accuracy and complexity of the predictive model, to compare estimated models. The model that minimizes AIC is preferable. This means choosing the simplest model (with fewer regressors) that explains the data well enough.

We followed the inverse elimination method, which involves beginning with including all variables in the model and removing one variable at a time, if the model thus obtained has a lower AIC. Finally, we obtained that the valid model with the lowest AIC is:

DOI: 10.24818/18423264/54.3.20.14

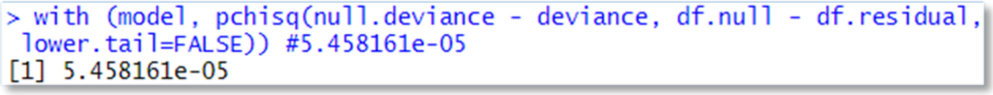


```
Call:
bayesglm(formula = seniority_level ~ max_depth_of_known_skills_in_graphs +
    first_known_skills_weight, family = "binomial", data = date,
    control = list(maxit = 100))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.99772  -0.47312  -0.02375   0.32242   1.59830

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -19.6399     7.7161  -2.545   0.0109 *
max_depth_of_known_skills_in_graphs   1.2808     0.6216   2.060   0.0394 *
first_known_skills_weight             0.4095     0.2338   1.752   0.0799 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.5256  on 19  degrees of freedom
Residual deviance:  7.8939  on 17  degrees of freedom
AIC: 13.894

Number of Fisher Scoring iterations: 44
```

In conclusion, the best logistic regression model, according to AIC, includes two explanatory variables: maximum depth of known skills and primary known skills weight. Both parameters of the model are statistically significant for an accepted 10% significance threshold. These two metrics are considered reliable in order to predict the seniority level of the candidate.

In addition to verifying the statistical significance of the model parameters there are other complementary tests for validating and measuring the model’s performance.

In order to decide on the model's creditworthiness, a statistical test is used based on model comparison with another model containing only a free term, called the null model. Test statistic is the difference between the errors of the two models and follows a χ^2 distribution (chi-squared distribution) with k (number of regressors) degrees of freedom.



```
> with (model, pchisq(null.deviance - deviance, df.null - df.residual,
 lower.tail=FALSE)) #5.458161e-05
[1] 5.458161e-05
```

The obtained p-value obtained of 0.0000545 is lower than 0.05 (accepted confidence threshold) and confirms that the model is significantly better than the null one.

In order to estimate the performance of the forecasting model, based on sample data, we used Receiver Operating Characteristics (ROC). The ROC plot represents on the OX axis the rate of candidates that were classified as senior but they were middle (false positive rate), and on the OY axis the rate of the candidates correctly classified as seniors (true positive rate). The obtained curve must be above the axis of the first bisector, so localized as much as possible to N-V, so that true positive rate is high and false positive rates is low. The area under the curve, usually referred to as AUC (Area under the ROC curve) is a measure of the model accuracy, so the closer it is to 1, the better the classifier.

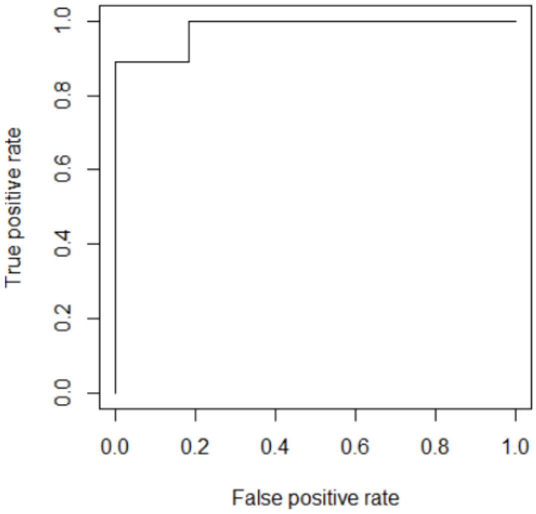


Figure 5. The ROC curve of the estimated model

In our case, as depicted in Figure 5, the area is 0.97, which suggests a very good performance of the model on how well separates the candidates from the analyzed group in those that are senior and those that are middle.

The previously validated model can be written as follows:

$$\text{logit(p)} = -19.6399 + 0.4095 * \text{primary known skills weight} + 1.2808 \\ * \text{maximum depth of known skills}$$

As in the case of linear regression, it is of interest to interpret the regression coefficients $\mathrm{b}_{\mathrm{i}=\overline{1,k}}$, which express the change in the logit quantity when $\mathrm{x}_{\mathrm{i}=\overline{1,k}}$increases with a unit. The sign of the coefficients indicates that there is a positive relationship between the response variable and the explanatory variables included in the model. An easier interpretation is given by e^{b_i} which express the effect of the regression coefficient on the success chances. In this sense, we applied in R the *exp()* function directly to the model coefficients and obtained the results presented in Table 3.

DOI: 10.24818/18423264/54.3.20.14

Table 3. Translating the model coefficients to success odds

Explanatory variable $(\mathbf{x_i})$	e^{b_i}
primary skills weight	1.506099
maximum depth of known skills	3.59934

We can conclude that:

- Increasing with a unit the primary known skills weight will increase the chances of a candidate to be senior by 50% or 1.5 times.
- Increasing with a unit the maximum depth of known skills will increase the chances of a candidate to be senior by 250% or 3.5 times.

7. Conclusions and future work

We combined semantic technologies with augmented reality in order to enhance the plain text from a CV content with context-aware information in regards to the technical background of the applicant. We developed the prototype of a CV screening system that helps the recruiter to upload the candidate's CV for a given position and displays a graph with the skills of the applicant and how those skills relate to others in the field. We tested our application on a CVs dataset and we collected a series of metrics we considered relevant for explaining the seniority level. We concluded that there is a positive relationship between both the maximum depth of known skills and primary known skills weight in the skill-set graph, and the seniority level of the candidate. We quantified also the effect those two exogenous variables have on the chances of an applicant to be senior.

Directions for future research include the extension of the proposed ontology with a large set of skills and their relations. The robustness and completeness of the modelled ontology is directly influencing the accuracy of the system that uses it. We also plan to refine the skills extractor component to ensure most of the technical competences encompassed in a CV are correctly identified.

Considering the new LinkedIn search paradigm, called Search by Ideal Candidates, proposed by V. Ha-Thuc et al. (2016), we plan to further improve our application and use the skill-set graph in order to compute the similarity between the candidates that apply for a job and other employees from the company, working on a similar position. This approach is known as item-to-item recommendation. Based on the computed similarity we will rank all the applicants and filter the ones that have a matching above a predefined threshold. In this manner the CV filtering process is fully automated and also takes into account not only the skills explicitly mentioned in the CV, but also the ones they are related with, providing a more trustworthy approach than manual selection performed by the HR personnel.

DOI: 10.24818/18423264/54.3.20.14

REFERENCES

[1] **AltexSoft (2018*),* *Software Engineer Qualification Levels: Junior, Middle, and Senior.*** Retrieved from https://www.altexsoft.com/blog/business/software-engineer-qualification-levels-junior-middle-and-senior/.

[2] **Bajenaru, L., Smeureanu, I. (2015), *An Ontology Based Approach For Modeling E-Learning in Healthcare Human Resource Management***; Economic Computation and Economic Cybernetics Studies and Research, vol. 49, no. 1, 2015, pp. 23-40; *ASE Publishing;*

[3] **Bloomberg D. (2014). *Building a Simple Text Recognizer.*** Retrieved from http://www.leptonica.org/.

[4] **Chen P., Peng Z. Li, D., Yang L. (2016)*, An Improved Augmented Reality System Based on AndAR***. Journal of Visual Communication and Image Representation, 37, pp. 63-69;

[5] **Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Roberts, A. (2014, November 17)*, Developing Language Processing Components with GATE Version 8.*** (The University of Sheffield, Department of Computer Science). Retrieved from https://gate.ac.uk/sale/tao/.

[6] **Dita F. (2016)*, A Foreign Language Learning Application using Mobile Augmented Reality*.** Informatica Economica, 20(4/2016), pp. 76-87;

[7] **Enachescu M.-I. (2019)*, Screening the Candidates in IT Field Based on Semantic Web Technologies: Automatic Extraction of Technical Competencies from Unstructured Resumes***, Informatica Economică vol. 23, no. 4/2019, pp. 51-65;

[8] **Ermilov T., Khalili A., Auer S. (2014)*, Ubiquitous Semantic Applications. International Journal on Semantic Web and Information Systems***, 10(1), pp. 66-99;

[9] **Gelder, A. B. (2012)*, Logistic Regression***, University of Iowa. Retrieved from: https://agelder.weebly.com/uploads/2/3/2/5/23257230/logistic_regression.pdf.;

[10] **Geroimenko V. (2012)*, Augmented Reality Technology and Art: The Analysis and Visualization of Evolving Conceptual Models***. 16th International Conference on Information Visualisation;

[11] **Ghosh H., Mallik A., Chaudhury S. (2015)*, Multimedia Ontology: Representation and Applications* (1st Ed.)**. CRC Press;

[12] **Grace-Martin K. (2018)*, The Difference between Logistic and Probit Regression. The analysis factor.*** Retrieved from https://www.theanalysisfactor.com/the-difference-between-logistic-and-probit-regression/;

[13] **Ha-Thuc V., Xu Y., Pradeep Kanduri S., Wu X., Dialani V., Yan Y., Gupta A., Sinha S. (2016)*, Search by Ideal Candidates: Next Generation of Talent Search at LinkedIn***, in WWW '16 Companion Proceedings of the 25th International Conference Companion on World Wide Web, Montréal, Québec, Canada, 2016, pp. 195-198;

[14] **Hahn E., Soyer R. (2005), *Probit and Logit Models: Differences in the Multivariate Realm*.** Retrieved from https://home.gwu.edu/~soyer/mv1h.pdfl

[15] **Han P., Zhao G. (2015)*, CAD-based 3D Objects Recognition in Monocular Images for Mobile Augmented Reality***. Computers & Graphics, 50, pp. 36-46;

DOI: 10.24818/18423264/54.3.20.14

[16] **Hitzler P., Krotzsch M., Rudolph S. (2010),** ***Foundations of Semantic Web Technologies***. Boca Raton (etc.): Chapman & Hall/CRC;
[17] **Jamali S., Shiratuddin M., Wong K., Oskam C. (2015),** ***Utilising Mobile-Augmented Reality for Learning Human Anatomy***. Procedia - Social and Behavioral Sciences, 197, pp. 659-668;
[18] **Karaman A., Erisik D., Incel O., Alptekin G. (2016),** ***Resource Usage Analysis of a Sensor-based Mobile Augmented Reality Application***. Procedia Computer Science, 83, 300-304;
[19] **Luo X. (2011),** ***The Cloud-Mobile Convergence Paradigm for Augmented Reality*** in Augmented Reality - Some Emerging Application Areas (Nee A.-Y.-C, Ed.). IntechOpen;
[20] **Majid N., Mohammed H., Sulaiman R. (2015),** ***Students' Perception of Mobile Augmented Reality Applications in Learning Computer Organization***. Procedia - Social and Behavioral Sciences, 176, pp.111-116;
[21] **Pocatilu P., Enachescu M. I., Dita A (2017),** ***Using Semantic Web Technologies for Augmented Reality based Mobile Applications***, The International Conference "Current Economic Trends in Emerging and Developing Countries, TIMTED 2017, Timisoara, 19-20 May 2017, http://www.timted.ro;
[22] **Sanchez J., Tello-Leal E., Carreon-Gutierrez J., Saldivar-Alonso V., Guerrero-Melendez T. (2013),** ***An Augmented Reality System Approach for Mobile Devices***. International Journal of Latest Research in Science and Technology, 2(5), pp. 9-11;
[23] **Schoonjans, F. (2019),** ***Logistic Regression***. Retrieved from https://www.medcalc.org/manual/logistic_regression.php
[24] **Sicilia M. (2014),** ***Handbook of Metadata, Semantics and Ontologies***. Singapore: World Scientific Pub. Co;
[25] **Sikos L. (2015),** ***Mastering Structured Data on the Semantic Web***. New York: Apress;
[26] **Thai, N. (2019),** ***Content Analysis with Apache Tika | Baeldung***. Retrieved from https://www.baeldung.com/apache-tika;
[27] **Weng E., Khan R., Adruce S., Bee O. (2013),** ***Objects Tracking from Natural Features in Mobile Augmented Reality***. Procedia - Social and Behavioral Sciences, 97, pp. 753-760.

DOI: 10.24818/18423264/54.3.20.14