sustainability

MDPI

Article

People Analytics of Semantic Web Human Resource Résumés for Sustainable Talent Acquisition

Sabina-Cristiana Necula * and Cătălin Strîmbei

Department of Accounting, Business Information Systems and Statistics, Faculty of Economics and Business Administration, Alexandru Ioan Cuza University of Iasi, 700505 Iași, Romania

* Correspondence: sabina.necula@uaic.ro

Received: 31 May 2019; Accepted: 23 June 2019; Published: 27 June 2019

Abstract: The purpose of this study was to define a data science architecture for talent acquisition. The approach was to propose analytics that derive data. The originality of this paper consists in proposing an architecture to work within the process of obtaining semantically enriched data by using data science and Semantic Web technologies. We applied the proposed architecture and developed a case study-based prototype that uses analytics techniques for résumé data integrated with Linked Data technologies. We conducted a case study to identify skills by applying classification via regression, k-nearest neighbors (k-NN), random forest, naïve Bayes, support vector machine, and decision tree algorithms to résumé data that we previously described with terms from publicly available ontologies. We labeled data from résumés using terms from existing human resource ontologies. The main contribution is the extraction of skills from résumés and the mining of data that was previously described with the Semantic Web.

Keywords: data science; talent management; Semantic Web; skills; analytics

1. Introduction

People analytics is fast becoming a key instrument in talent management. Human resource analytics, also called talent analytics, is the application of considerable data mining and business analytics techniques to human resources data [1,2]. A key aspect of people analytics is represented by data about people or human resources. In recent years, there has been an increasing interest in data science and analytics on data. Evidence [3–6] suggests that data science supports organizations by providing descriptive, predictive, and prescriptive analytics. Talent management could benefit from all these techniques, especially in the phase of talent acquisition. Talent acquisition is an integral part of talent management. Unfortunately, there is limited literature in human resource analytics to guide the use of machine learning algorithms [7]. Even if the skills and ability to conduct these analyses are present, it is still a challenge to gather the data necessary to turn information into results [8].

To date there has been little agreement on the necessary data for talent analytics, but there is a common agreement that skills, work experience, and education form the basis of building a résumé. Websites like LinkedIn, Indeed, Jobup, and others try to achieve better matching between job positions and résumés. LinkedIn applies, for example, machine learning to individual profiles, and extracts features like skills, seniority, and industry. Similar features are extracted from the content on the job listing. Furthermore, logistic regression models are used to rank relevant jobs for a given member using these features [9].

Firstly, we analyzed the literature existent in human resource analytics. Secondly, we studied the subject from the perspective of the Semantic Web. In the last decade, the nonparametric methods (machine learning algorithms) have gained great attention in human resource management practice field. Examples of the use of analytics in talent management are data mining (extracting and

examining data from large databases), sentiment analysis, and controlled tests such as A/B testing [10]. However, despite the benefits of using and implementing these technologies, little is known about how to benefit from the Semantic Web and analytics on data, specifically about how to link and derive data from people's résumés. In this study, we proposed a Semantic Web data science architecture and validated it on résumés described with the Semantic Web.

Srivastava et al. [11] provided several predictive analytics to address talent acquisition needs such as predicting joining delay, selection likelihood, and offer acceptance likelihood. Dutta et al. [12] used data mining for getting insights and text mining for talent acquisition efficiency improvement. Faliagka et al. [13,14] proposed a system that implements candidate ranking, using objective criteria that are made available from the applicant's LinkedIn profile. The candidate's personality features are also extracted from their social activity using linguistic analysis. Faliagka et al. [14] used text mining of LinkedIn for creating profiles and linguistic analysis for inferring personality characteristics. Palshikar et al. [15] extracted attributes from candidate résumés while planning to combine information from multiple online and social platforms for the technical and domain skills using extraction tools. Mooney and Bunescu [16] applied knowledge extraction from unstructured text using text mining. With increased use of machine learning and natural language processing techniques, Téllez-Valero et al. [17] and other researchers tried to solve this problem of automatic extraction. With résumés, different extraction techniques are used to make the candidate selection process [18] easier and more automatic.

Previous studies [19] reported a machine learning application for the human resource data mining problem. Xie and Tang [20] used fuzzy neural networks for human resource. With respect to recruitment data mining, there are studies that use clustering and classification algorithms [20] to prove that fuzzy C-means and K-means clustering techniques are not suitable for this type of data distribution. It has been observed that trees constructed with the C4.5 algorithm (decision tree algorithm) have better accuracies. Another type of application is that of profile development [21].

Aldarra and Munoz [22] applied J48 algorithm to construct a Linked Data-based decision tree classifier to review movies. They used the SPARQL Protocol and RDF Query Language (SPARQL) queries to derive features. Mehenni and Moussaoui [23] built a regression model for predicting the most useful links that will be connected to build a multi-relational decision tree for heterogeneous databases. Sanchez-Marono et al. [24] discussed the use of decision trees learned from questionnaire data as behavioral models for the agents comparing various pre-processing methods and exploring their differences.

There is very little scientific understanding of skills from résumés. In addition, to the best of our knowledge, only a limited number of research papers comparing and evaluating the performance of different analytics algorithms with different training sample strategies using résumé data have been published.

Current implementations of Linked Data mining are promising [18,25,26]. However, the full potential of the Semantic Web and Linked Open Data for data mining and knowledge database discovery is still to be unlocked [27].

Now, there are some developments of human resource ontologies, such as the Human Resources Management Ontology [28]. The literature notes that the current concerns are [29]: (1) publishing job postings and applicant profiles enriched through domain ontologies/controlled vocabularies, (2) pre-selection of the candidates based on semantic matching techniques implemented in addition to these ontologies and the associated automated reasoning, and (3) delivering interview recommendations to employers or suitable open positions to job seekers based on the semantic matching of the annotated applicant profiles with the job postings [30]. Our work addresses applicant skills profiles enriched through domain ontologies and analytics in the context of data analytics Semantic Web architecture.

The major objective of this study was to investigate the possibilities that data analytics offers for skills identification. The research goal of this paper is to increase the efficiency of the analytical data processing through semantically described data and query processing. The term efficiency, in this

paper, is understood in a broader sense: by semantically describing data, the analytical processing is improved.

Methodologically, the research presented in this article follows the design science research paradigm [31]. It uses data from a case study to define its solution objectives. The artifact, the information system architecture, is constructed based on the analogy to the human resource process, the literature, and data from the case study. To evaluate the information system architecture, a prototype is developed and its limitations are analyzed through accuracy measures.

The case study takes into consideration résumé data described with ontologies. We identified the candidates' skills by discovering relationships between the employee's skills, work experience, and education on one side, and the current position held on the other side. The structure of this paper is as follows: Section 2, Materials and Methods, presents the research methodology, Resource Description Framework (RDF) knowledge base construction, and feature engineering, Section 3 discusses the results, and Section 4 presents the discussions.

2. Materials and Methods

The "validation in context" is a key feature. Therefore, we first proposed a Semantic Web data science architecture and, after deciding that the proper context is résumés websites, we validated the artifact.

Our article proposes analytics in an architecture that also includes Semantic Web technologies with the specific objective of identifying and quantifying the contribution of these technologies, starting from the idea that potential relationships established on the semantic basis can contribute to the analytical model of data.

2.1. Research Hypothesis

Our work is guided by the following research questions: (1) Is it possible to obtain semantically improved data by using data science on Semantic Web-described data? and (2) What will the necessary architecture to support data science on Semantic Web data look like?

We hypothesized that for the process of linking résumés data a Semantic Web data science architecture can be established. This design is based on a set of architectural decisions made to discover links between Linked Data, specifically between résumés data described with ontologies terms (H1). We tested this architecture on a case study based prototype. We validated the results by analyzing the accuracies, receiver operating characteristic (ROC) and precision recall curve (PRC) values of different classification algorithms applied on the dataset and on the dataset that we enriched with features obtained by aggregating data.

We further hypothesized that using this architecture discovers links between data (H2). We derived links between data by discovering the best predictors for every skill. We validated the results by analyzing the accuracies and ROC and PRC values of the decision tree algorithm applied on the dataset that we enriched with features obtained by aggregating data.

Figure 1 presents the structure of a résumé. Each résumé contains information about work experience (responsibilities and position held), education, and additional information such as technical skills.

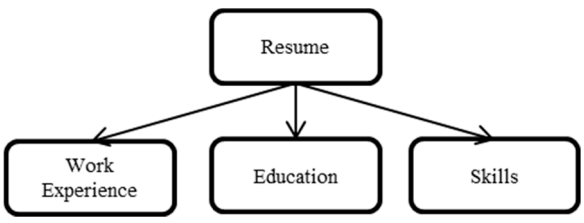

Figure 1. The structure of a résumé from Indeed résumé website.

The main idea was to structure résumés on semantic basis. Figure 2 provides an overview of the architecture as it is currently implemented. The components of the architecture arise from the corresponding architectural decisions made for pragmatic, technical and scientific reasons.

(1) The web scraper component seeks résumé data across the Indeed résumé website [30]. It extracts data from résumés written in HTML and saves data in the comma-separated-value (CSV) format.
(2) The mapping engine integrates data published using different vocabularies from the human resource ontology published by the Ontology Engineering Group. It transforms data from CSV to RDF by using terms defined as classes, sub-classes, and properties from other RDF files that represent ontologies.
(3) The résumé RDF processor labels different features of the data mining classifier model. It uses SPARQL to query data from RDF and derives the features.
(4) The classifier models use data and derive the prediction rules.

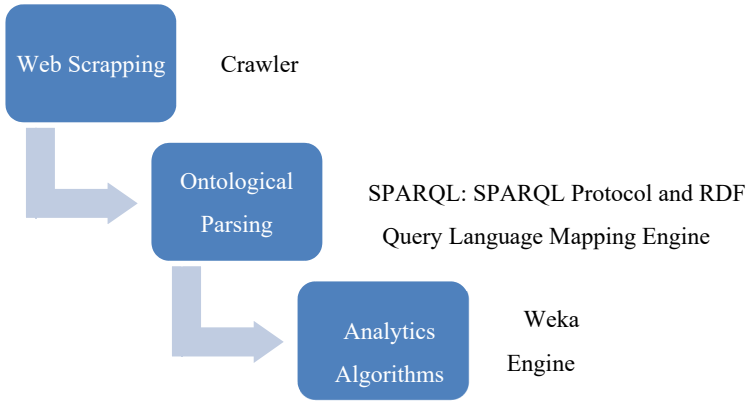

Figure 2. The architecture. Source: Our own projection.

The source code is available at https://github.com/catalinstrimbei/rdf-mining-hr [32].

2.2. Data Acquisition

Currently, there are no public linked datasets that contain human resource data, about competences or résumé data. Therefore, we obtained the dataset to build our classifiers by creating a web scraper on the Indeed résumé web site (https://www.indeed.com/resumes) and transformed the data into Turtle/Resource Description Framework by using OpenRefine [33]. We scraped data from Indeed, a website that contains data about people résumés publicly accessible in html online format. Résumés' acquiring must relate to a keyword to search for résumés. We intended to obtain résumés for people belonging to the same field of work. Therefore, we limited our searches to résumés related to the Java industry keyword. We used a word cloud to identify the main keywords encountered in the skills section from every résumé. This way we extracted 677 web addresses that link to résumés from the Java industry. We parsed these HTML pages and extracted information from 213 résumés. Specifically, only 213 résumés from the 677 résumés presented information structured according to Figure 2. We processed data and initially stored it in comma-separated-value (CSV).

The résumés data had to be transformed in RDF according to public ontologies found in the human resource field. Therefore, we mapped data from résumés to ontology's concepts and properties according to our own human resource ontology that extends the human resource ontology published by Ontology Engineering Group (OEG). The aim of this ontology is to represent knowledge related to the human resource hiring process. The human resource ontology developed by the Ontology Engineering Group is suitable for our purpose and is available at http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/ontologies/99-hrmontology/ [34].

We used the following ontologies: JobSeeker, Occupation, Education, Competence, and Skill. The public URIs do not work and therefore we adapted the namespaces of every ontology file. We also combined the JobSeeker, Occupation, and Education ontologies into a single ontology, (e.g., JobSeeker) because the fine granularity of Education and Occupation did not present findings of interest with respect to the research scope. The focus is on work experience and skills. The knowledge base consists of 213 résumés with data described by the JobSeeker, Competence, and Skill ontologies, which are accessed using SPARQL queries to generate our set of features to train the classifiers.

Figure 3 presents the key classes and properties of the ontologies used. A job seeker has work experience, candidacy, education, competence, and skill as a subclass of competence. Central to the ontology is the WorkExperience concept and its object properties that relate the concept of Work Experience to Candidacy and JobSeeker. The Candidacy concept requires Competence, and Skill is a subclass of Competence.

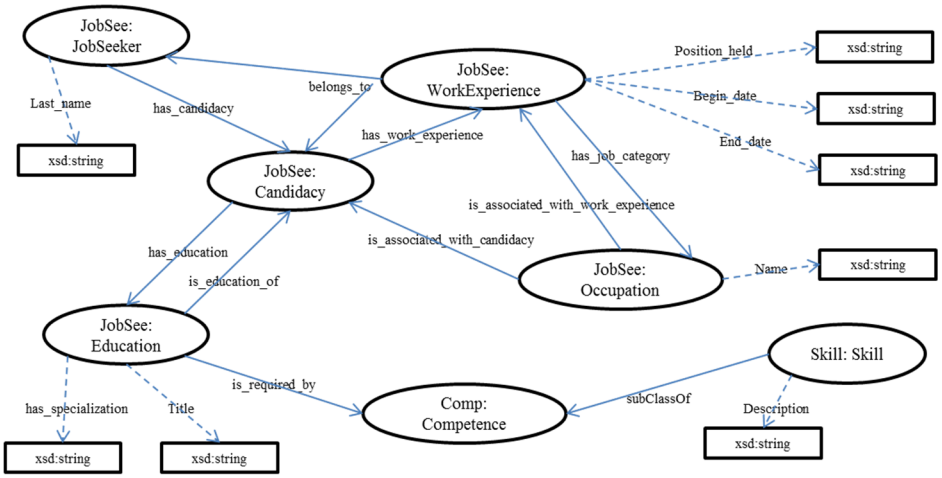

Figure 3. Key classes and properties of the ontology. Source: Our own projection.

The data properties (depicted with dashed arrows in Figure 2) allow the values of Work Experience, Education, and JobSeeker to be specified.

Every resource in our dataset is identified using a uniform resource identifier (URI). URIs have been designed with simplicity and manageability principles in mind.

The matching between a résumé and the terms from ontologies is presented in Figure 4.

The major purpose of data processing was to access data about work experience, education and skills. Besides storing raw data and transforming it to RDF, we also derived some new features, like the total years of experience, the years of experience in the current job position, and the average experience (measured in years) in every position held. These features were derived by using SPARQL. Operationalization of variables is described in Table 1.

For processing Semantic Web data, we identified two different types of features: (1) features derived with SPARQL and (2) features derived with SPARQL by aggregating data.

The information concerning skills splits between different technical skills. We restricted the skills that presented interest at SOA, NoSQL, SQL, Java, Java Web, and Java Persistence. We derived these categories by using a word cloud. Therefore, we queried the RDF data through SPARQL queries to find out which candidates have these kind of skills.

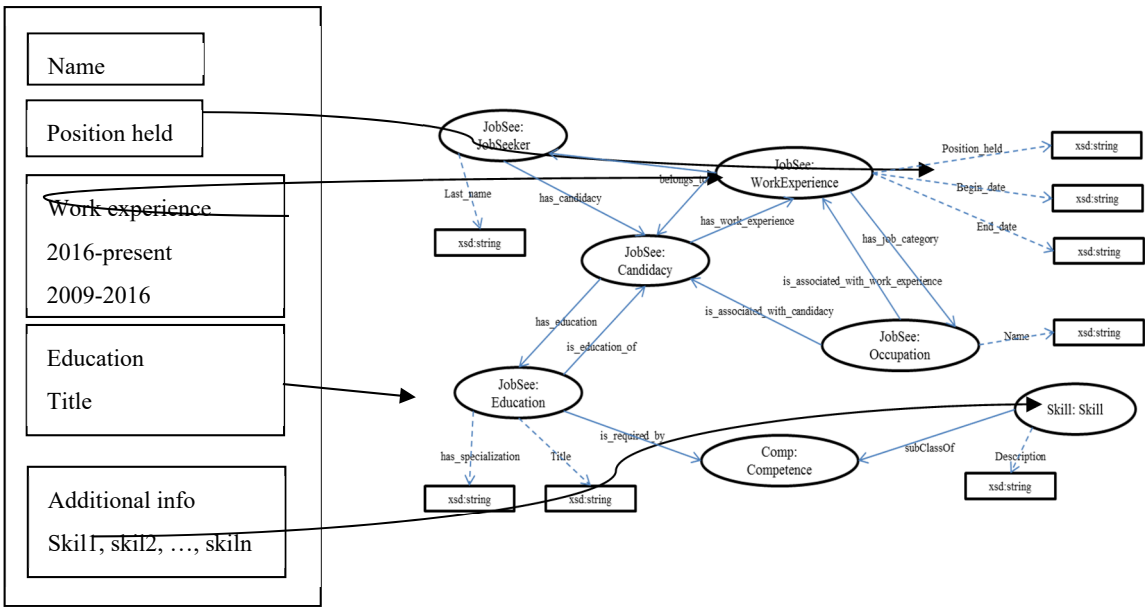

Figure 4. Indeed résumé match with ontologies concepts. Source: Our own projection.

Table 1. The operationalization of variables.

Feature	Ideas	Details
Total years of experience [35]	Extensive experience of activities in a domain is necessary to reach very high levels of performance.	"Expert performance is acquired gradually and the effective improvement of performance requires the opportunity to find suitable training tasks that the performer can master sequentially"
The years of experience at the current job position [36]	Values that are too big or too small are subject to further analysis	"Job satisfaction is positively correlated with mission valence, commitment, person–job fit, flexible work, pay, innovation, and a variety of other individual and organizational factors"
The average of the years of experience in every position held [37]	Variety in work experiences might influence forming high performers	"According to 50 senior executive search professionals the study surveyed, the average executive today will work in five companies; in another 10 years, it might be seven". "Ineffective people often stay in position for years".
The total number of positions held (Position_count) [38]	The total number of positions might influence forming high or low performers	"Job satisfaction more strongly determines organizational performance than organizational performance determines job satisfaction"

Source: Our own projection.

The information related to work experience splits in different positions held and their corresponding beginning and ending dates. We queried the data through SPARQL queries to obtain the last position held, the total years of experience and the average time spent at one position. In addition, we queried the data to find out the total number of positions held.

The information concerning education splits between the education level and the corresponding specialization. We identified computer science, information technology, electronics, computer engineering, software engineering, computer applications, and other specializations that we grouped as OTHER. We queried the data to obtain the education level (master or bachelor) and the corresponding specialization field. We derived 15 features extracted from data using SPARQL queries. The next section presents the obtained results.

Features defining and SPARQL specifications are presented in Table 2.

In order to obtain the current position held, we created a SPARQL query that extracts the positions held by each job seeker. After querying the data, the provided sample consists of 120 résumés

along with the anonymized URIs, objective (position held), work experience, education, and skills. Our resulting RDF knowledge base comprises 14,206 RDF triples.

Table 2. Features and their corresponding SELECT clauses from SPARQL queries.

Features (Attributes)	Types/Values	SPARQL SELECT Clauses
Job_Seeker	String	
Education_Title (f1)		Master's, Bachelor's, Other
Education_Title_Spec (f2)	Computer science, information technology, electronics, computer engineering, software engineering, computer applications, OTHER	Computer science, information technology, electronics, computer engineering, software engineering Master's, Bachelor's, Other, computer applications, OTHER
Java_Programming_Skills (f3)	True, False	Java, JEE, JSE, J2SE, J2EE
SQL_Programming_Skills (f4)	True, False	SQL, Oracle
NOSQL_Programming_Skills (f5)	True, False	NOSQL, Mongo DB
UML_Skills (f6)	True, False	UML
SOA_Developer_Skills (f7)	True, False	Web Service, SOA, REST, SOAP, JAX-RS, JAX-WS
Java_Web_Developer (f8)	True, False	Servlet, JSP, JSF, Struts
Web_Developer_Skills (f9)	True, False	HTML, Javascript, JQuery, Angular
DB_Developer_Skills (f10)	True, False	SQL, Oracle, MySQL, Postgres
Java_Persistence_Skills (f11)	True, False	JDBC, JPA, Java Persistence API, Hibernate
Years_Experience_last (f12)	Numeric	MAX(end_date)-MAX(begin_date)
Position_count (f13)	Numeric	COUNT(Position_held)
Years_experience_position (f14)	Numeric	AVERAGE(MAX(end_date) -MIN(begin_date), Position_count)
Position_held (f15)	JAVA developer, JEE developer, software engineer, other developer, other programmer, analyst, other engineer, other	JAVA developer, JEE developer, software engineer, other developer, other programmer, analyst, other engineer, other

Source: Our own projection.

Our features contain mixed continuous (numerical) and dichotomous (categorical) types that can be handled by the data mining techniques. Figure 5 summarizes the features used in this work.

Concerning the size of the sample, we studied the learning curve (Figure 6). Learning curves are a tool to do a quick check on the models at every point in the machine learning workflow. Bias and variance are inherent properties of estimators and we usually have to select learning algorithms and hyper parameters so that both bias and variance are as low as possible. In order to study the learning curve, we applied regression on our data. For regression, the perfect scenario is when both curves converge toward an MSE of 0.

The training data is fitted very well by the estimated model. If the model fits the training data very well, it means it has low bias with respect to that set of data. Therefore, we decided that the sample size is proper. We consider that the sample size is sufficient for generalization, as replication with other populations or conditions helps to define parameters related to education, work experience and skills.

To date, various methods have been developed and introduced to mine data.

We used the k-nearest neighbors algorithm (k-NN), naive Bayes classifiers, support-vector machine, random forest, regression, and the decision trees technique. The C4.5 classifier, a well-liked tree based

classifier, is used to generate a decision tree from a set of training examples. Nowadays C4.5 is renamed as J48 classifier in WEKA tool, which is an open source data mining tool. The heuristic function used in this classifier is based on the concept of information entropy [39]. We used WEKA to build our classifiers. To test the algorithms we choose 10-folds cross-validation method.

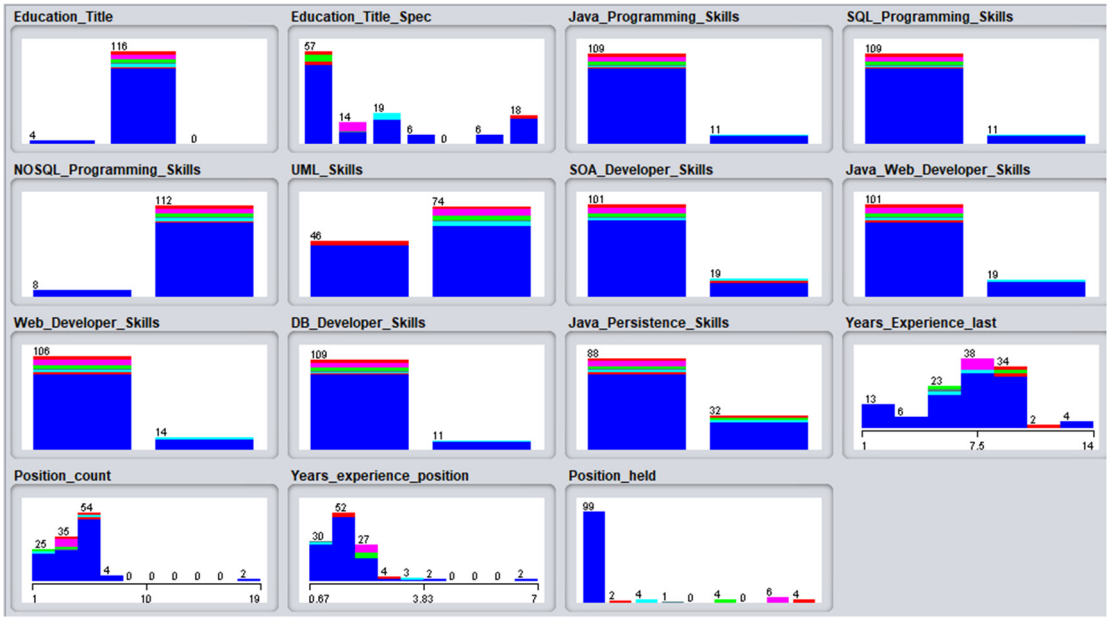

Figure 5. Visualization of attributes. Source: Weka visualization of attributes.

Figure 6. The learning curve. Source: Our own projection realized in Python on our dataset.

In our study, we chose the J48 algorithm to construct the decision tree. J48 implementation is widely used in research papers [40].

3. Results

The proposed task was to predict skills that a job seeker has.

We applied different algorithms on data. The accuracies, ROC, and PRC values for every algorithm are presented in Table 3.

In order to study the performance of the algorithms, we presented also the receiver operator characteristic (ROC) (Figure 7) and the precision recall curve (PRC) values. Davis and Goadrich [41] studied ROC and PRC. They explained that for skewed datasets the PRC values are more informative than ROC values. An optimal classifier will have ROC and PRC area values approaching 1, with 0.5 being comparable to random guessing.

Table 3. Algorithms and their accuracies. ROC: receiver operating characteristic; PRC: precision recall curve; k-NN: k-nearest neighbor.

Algorithm	Accuracy for the Dataset Containing Aggregated Features	ROC	PRC	Accuracy for the Dataset not Containing Aggregated Features	ROC	PRC
Classification via Regression	0.94	0.836	0.950	0.86	0.667	0.888
Support vector machine	0.86	0.661	0.875	0.85	0.696	0.887
k-NN	0.95	0.936	0.983	0.87	0.907	0.979
Naïve Bayes	0.84	0.784	0.941	0.76	0.573	0.863
Random forest	0.98	0.998	1.000	0.87	0.904	0.979
Decision tree	0.80	0.822	0.953	0.82	0.474	0.818

Source: Our own analysis realized with Weka.

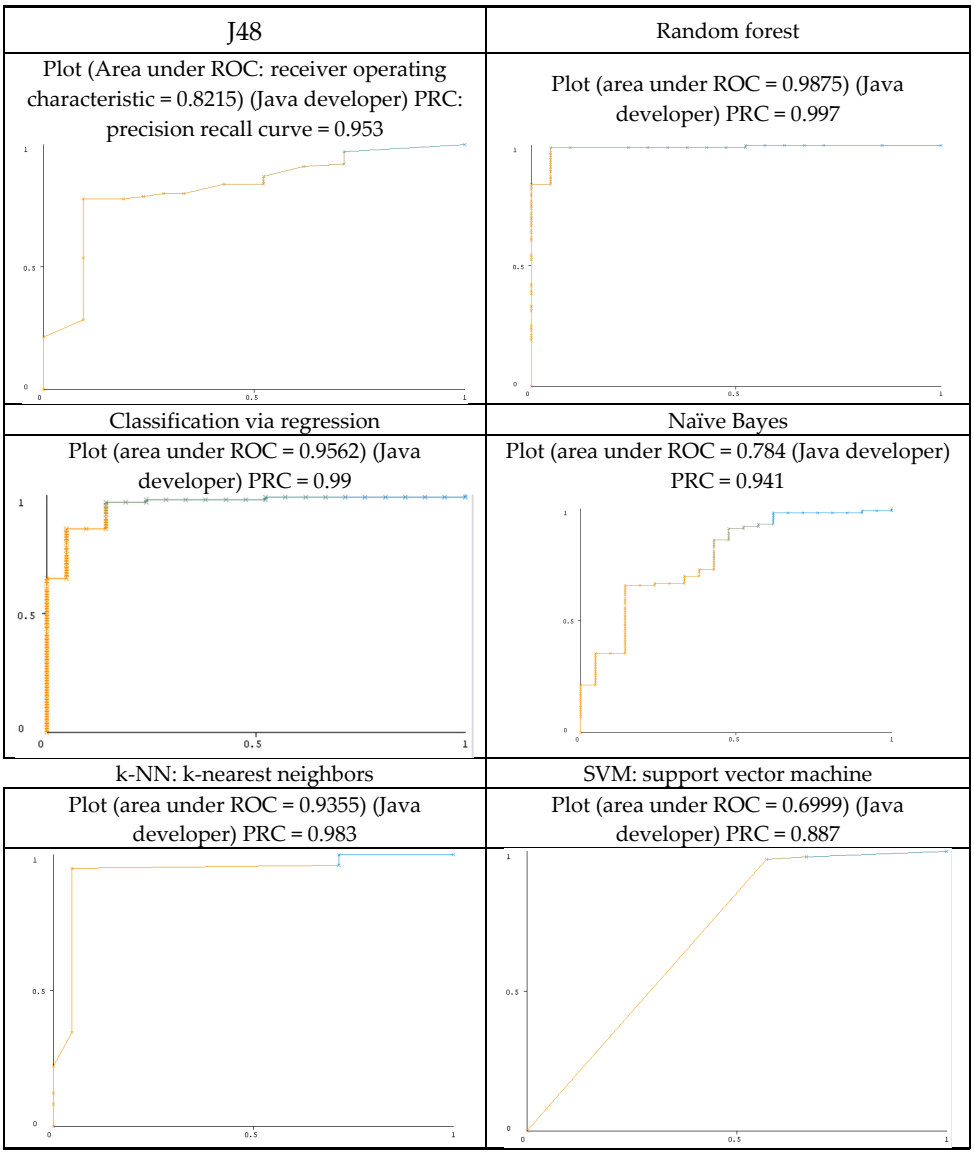

Figure 7. ROC and PRC of different algorithms for the Java-developer class (with aggregate features) Source: Our own projection realized with Weka.

It is important to notice that the accuracy of the decision tree applied on the features that do not include aggregations is greater than the accuracy of the decision tree applied on the features that include aggregations with 0.02. This is the only algorithm that presented this behavior of accuracies. We also mention that the ROC value is only 0.474 and the PRC is 0.818; therefore, we concluded that the classifiers models performed better on data enriched with features obtained through aggregations.

Java-developer class has the highest number of instances. We noticed that for Java developer class J48, k-NN and random forest have good ROC curves.

For the J48 algorithm, we used a pruned tree, meaning that we tried to avoid overfitting. In addition, we used binary split for nominal attributes. We applied J48 classification algorithm, by using a pruned tree method, binarySplits, on nominal attributes, a 0.25 confidence factor, and a 10-fold cross validation method for testing the model. Table 4 presents the values for J48.

Table 4. Precision, Recall and the F-Measure for the J48.

	Precision	**Recall**	**F-Measure**
Java developer	0.858	0.919	0.888
JEE developer	0	0	00
Software engineer	0	0	0
Other—Developer	0	0	0
Other—Programmer	0	0	0
Analyst	0	0	0
Other—Engineer	0.857	1.0	0.923
Other	0	0	0

Note: The features of the vector are represented by all the attributes, except for Job_Seeker. The class attribute/target is Position_held Source: Our own projection realized with Weka.

The confusion matrix of the J48 classifier is presented in Table 5.

Table 5. The confusion matrix for the Position_held J48 classifier.

	a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**Accuracy**
a = Java developer	91	0	4	0	0	1	1	2	97/120 = 80.83
b = JEE developer	2	0	0	0	0	0	0	0	
c = Software engineer	4	0	0	0	0	0	0	0	
d = Other—Developer	1	0	0	0	0	0	0	0	
e = Other—Programmer	0	0	0	0	0	0	0	0	
f = Analyst	3	0	0	0	0	0	0	1	
g = Other—Engineer	0	0	0	0	0	0	6	0	
h = Other	4	0	0	0	0	0	0	0	

Source: Our own projection realized with Weka.

It can be noticed that the model has a good value of prediction, so we proceeded to analyze further.

We observed that the attribute splitting the data is Years_experience_position. Moreover, the Java developers, the class that has the highest number of correctly classified instances is determined by being skilled in Java programming and SOA. Therefore, in order to determine the related skills, there is the need to analyze data for every skill.

We started to build the decision trees for every skill. Table 6 presents the accuracy details for every skill classifier.

Table 6. Detailed accuracy metrics for every skill (ROC: receiver operating characteristic, PRC: precision recall curve).

Precision	Recall	F-Measure	ROC	PRC	Class
Java_Programming_Skills (output) = (f1, f2, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15)					
1.0	1.0	1.0	1.0	1.0	True
1.0	1.0	1.0	1.0	1.0	False
SQL_Programming_Skills (output) = (f1, f2, f3, f4, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15)					
1.0	1.0	1.0	1.0	1.0	True
1.0	1.0	1.0	1.0	1.0	False
NOSQL_Programming_Skills (output) = (f1, f2, f3, f4, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15)					
0.806	0.543	0.649	0.804	0.747	True
0.764	0.919	0.834	0.804	0.845	False
SOA_Developer_Skills (output) = (f1, f2, f3, f4, f5, f6, f8, f9, f10, f11, f12, f13, f14, f15)					
0.962	1.0	0.981	0.932	0.976	True
1.0	0.789	0.882	0.932	0.872	False
Java_Web_Developer_Skills (output) = (f1, f2, f3, f4, f5, f6, f7, f9, f10, f11, f12, f13, f14, f15)					
0.980	0.960	0.970	0.954	0.948	True
0.810	0.895	0.850	0.954	0.803	False
Web_Developer_Skills (output) = (f1, f2, f3, f4, f5, f6, f7, f8, f10, f11, f12, f13, f14, f15)					
0.972	1.0	0.986	0.893	0.972	True
1.0	0.766	0.880	0.893	0.811	False
DB_Developer_Skills (output) = (f1, f2, f3, f4, f5, f6, f7, f8, f9, f11, f12, f13, f14, f15)					
1.0	1.0	1.0	1.0	1.0	True
1.0	1.0	1.0	1.0	1.0	False
Java_Persistence_Skills (output) = (f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f12, f13, f14, f15)					
0.863	0.932	0.896	0.793	0.883	True
0.760	0.594	0.667	0.793	0.670	False

Source: Our own projection realized with Weka.

J48 predicts good models with high accuracy for every identified skill. We observe that for Java programming skills, SQL programming skills, and database developer skills the accuracy is the highest. In fact, the classifiers predicted that those who have Java programming skills have SQL programming skills and vice versa. Those that have Java programming skills have also database skills.

J48 predicted better than the baseline models for the rest of the skills. The J48 classifier built for Java programming skills identified that SQL programming skills is the node that splits instances. In addition, by running different tests, we found out that a good predictor for SQL programming skills and for database programming skills is Java programming skills.

The Java web programming skills J48 classifier (Figure 8), identified that Java programming skills, Java persistence skills and NOSQL programming skills are the features that split the data.

The NOSQL programming skills J48 classifier (Figure 9) identified that Java persistence skills, UML skills and Java web developer skills are features related to skills that split the data.

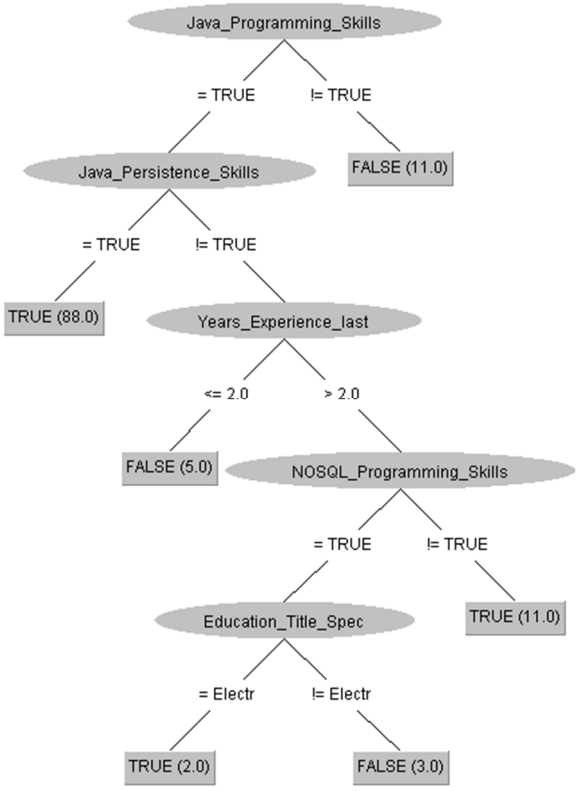

Figure 8. Java web developer skills. Source: Our own projection realized with Weka.

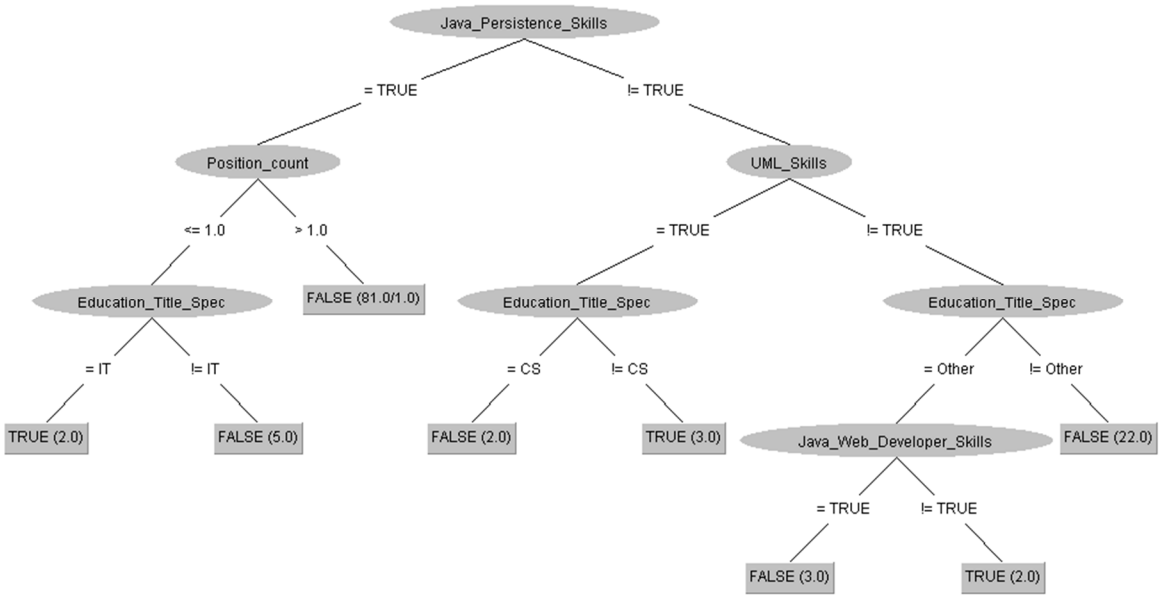

Figure 9. NOSQL programming skills decision tree. Source: Our own projection realized with Weka.

The SOA programming skills J48 classifier (Figure 10) identified that Java programming skills are related to SOA. In addition, it seems that experience at job positions that are not designed for JEE developers is important in classifying instances as having SOA programming skills.

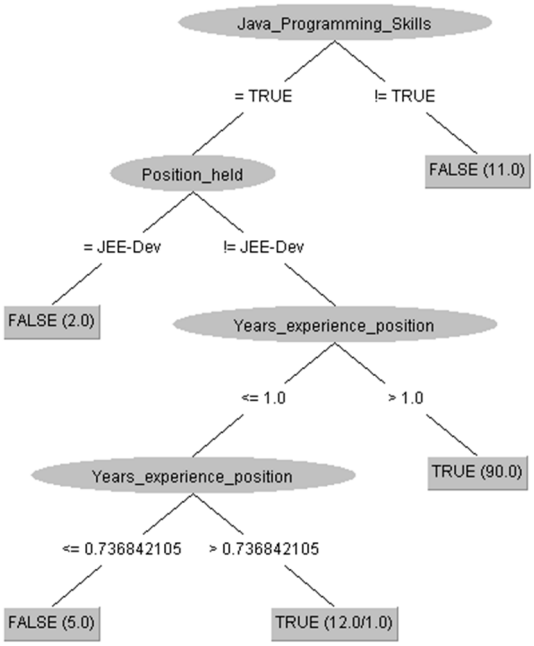

Figure 10. SOA decision tree. Source: Our own projection realized with Weka.

These findings further support the idea of using ontologies to better describe data from people résumé data.

4. Discussion

As we mentioned in the literature review, employees' skills are of great concern to talent management. A rich body of the literature has focused on the importance of competences for business, but little is known about how to identify skills that employees have starting from data presented in job seekers' profiles. The exploration and development of new skills, career paths and education levels require scientists and human resource experts to extract knowledge from multiple sources of information.

This study contributes to the existing literature by specifically analyzing how to discover links between data by using Semantic Web technologies and analytics.

Mining data from people résumés brings to surface relations between résumés data and employability. Our approach states that starting from the relation between work experience, education, and skills on one side and position held on the other side it is possible to derive links between skills. This approach uses data about skills represented in a Linked Data structured representation format.

We compared the results with the findings of previous work. The main contributions of our work presented in relationship with other studies results are surveyed in Table 7.

Our current study found that the J48 classifier built for Position_held identified that the specialization chosen for studying, UML skills, Java programming skills, the number of years spent in average for every position held and the number of years spent at the last position held are the features that split the data. We built the classifier by using binary split for nominal features.

The most interesting finding was that when we built the same classifier without binary splitting the data of nominal features, we observed that the features that split the data are: number of years spent in average at every position held, Java programming skills, SOA programming skills, the number of positions held, and the number of years spent at the last position held.

Table 7. The main contributions of our work. (RDF: Resource Description Framework, k-NN: k-nearest neighbors, SPARQL: SPARQL Protocol and RDF Query Language, NOSQL: not only SQL)

Important for Analytics		
Contribution	**Our Architecture**	**Others**
Describing human resource data with RDF	Yes	Yes, but not for résumé data. There are many studies that describe data with RDF [42], propose tools to automatically describe data [43], or publish RDF data [44], but not résumé data.
Features derived with SPARQL by using aggregate functions	Yes In the case study, we defined the number of years spent in average at every position held, the number of positions held, and the number of years spent at the last position held. We found that they are important for Position_held and for predicting better some skills	Partial aggregate functions but to movie reviews [22].
Features derived from Linked data mining	Yes In our case study, the Java web programming skills J48 classifier (Figure 9) identified that Java programming skills, Java persistence skills, and NOSQL programming skills are the features that split the data. The NOSQL programming skills J48 classifier (Figure 10) identified that Java persistence skills, UML skills, and Java web developer skills are features related to skills that split the data.	Partial [11] use of linear regression to predict human resource employability, but no classification machine learning algorithm to derive skills. In addition, they use linguistic analytics. We used the Semantic Web in order to offer a better representation. This representation is useful in future dataset querying so that when searching the dataset for people with certain skills, enhanced information be provided. Mochol et al. [45] use a dictionary of synonyms, but not the Semantic Web. Kessler et al. [46] classify the applicants with the support vector machine algorithm. We used decision tree algorithm to find the best predictors. To predict position held, we applied random forest, classification via regression, naive Bayes, k-NN, support vector machine and decision tree.

Source: Our own projection.

Table 8 shows examples of the three types of features that our study analyzed with the aim of better describing the results of the paper.

Table 8. Types of feature examples.

Features derived with SPARQL queries
JobSeeker has_average_work_experience xsd:integer
Features derived from other features (skills related to data from people résumé)
Java_Web_Developer_Skills is_related_to SOA_programming_skills
Attribute selection by using data mining
SOA_programming_skills isImportant in hiring Java developers

Source: Our own projection.

In this study, all classification algorithms performed better when they were applied on the dataset previously enriched with features resulted from aggregating data, like: years of experience at the last position, the total number of position held, the average years of experience on each position. The accuracies, ROC and PRC values proved that predicting the position held is improved when using features derived by aggregations.

Furthermore, describing data in terms belonging to ontologies provided the possibility to derive links between skills. Starting from a compact description of skills, we queried with SPARQL each skill, we labeled with terms from ontologies for future analysis. The results proved that different skills are determined by the existence of other skills. In addition, the features derived by aggregations are predictors for some skills, becoming normal to infer that some skills come with experience or by implying in diverse activities during changing jobs.

Taken together, these results suggest that linking terms and properties from diverse vocabularies help inferencing on data belonging to eRecruitment websites.

In this study, we described an approach for automatically detecting the important attributes for the process of hiring job seekers. Our approach is able to detect three types of features. We described the method used in our approach: (1) using the Indeed résumé database, (2) describing data with concepts belonging to ontologies, (3) querying data with SPARQL, (4) defining features, and (5) mining data with analytics.

The main contributions of our work are as follows. Firstly, we proposed analytics for discovering the important features for hiring job seekers starting from résumé data. The method operates by selecting attributes with a high information gain ratio. The attributes were previously defined with SPARQL queries. Secondly, an experimental analysis was conducted on Indeed's résumé data with the aim of applying the method.

This study has shown that analytics on features derived with Semantic Web technologies help identify better links between data. Furthermore, identifying which skills determine other skills at the level of an entire population also has a large impact on the data analysis.

Organizations tend to select project teams based on experience, availability, and past individual performance. One future application would be to predict the success rate of a team based on team composition and context variables.

Finally, we mention the impact of using the results of our work on graph database from the LinkedIn website together with other personal job seeker web pages or other job website portals. We believe that the granularity of skills' descriptions will impact on the results of analytics.

Author Contributions: S.-C.N. and C.S. contributed equally.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Carlson, K.D.; Kavanagh, M.J. HR metrics and workforce analytics. In *Human Resource Information Systems: Basics Applications and Future Directions*, 1st ed.; Kavanagh, M.J., Thite, M., Eds.; Sage Publishing: Thousand Oaks, CA, USA, 2012; pp. 150–174.
2. Bányai, T.; Landschützer, C.; Bányai, Á. Markov-chain simulation-based analysis of human resource structure: How staff deployment and staffing affect sustainable human resource strategy. *Sustainability* **2018**, *10*, 3692. [CrossRef]
3. Lunsford, D.L. An Output Model for Human Resource Development Analytics. *Perform. Improv. Q.* **2019**, *32*, 13–35. Available online: https://onlinelibrary.wiley.com/doi/abs/10.1002/piq.21284 (accessed on 15 June 2019). [CrossRef]
4. Fotache, M. Data Processing Languages for Business Intelligence. SQL vs. R. *Inform. Econ.* **2016**, *20*, 48–61. Available online: http://revistaie.ase.ro/content/77/05%20-%20Fotache.pdf (accessed on 15 June 2019). [CrossRef]
5. Păvăloaia, V.D.; Georgescu, M.R.; Popescul, D.; Radu, L.D. ESD for Public Administration: An Essential challenge for inventing the future of our society. *Sustainability* **2019**, *11*, 880. [CrossRef]
6. Minastireanu, E.A.; Mesnita, G. Light GBM Machine Learning Algorithm to Online Click Fraud Detection. *J. Inform. Assur. Cybersecur.* **2019**, *2019*. Available online: https://ibimapublishing.com/articles/JIACS/2019/263928/ (accessed on 15 June 2019).
7. King, K.G. Data Analytics in Human Resources: A Case Study and Critical Review. *Hum. Resour. Dev. Revi.* **2016**, *15*, 487–495. Available online: https://journals.sagepub.com/doi/abs/10.1177/1534484316675818 (accessed on 15 June 2019). [CrossRef]
8. Fitz-Enz, J.; Mattox, J. *Predictive Analytics for Human Resources*; John Wiley: Hoboken, NJ, USA, 2014.
9. Purohit, S.R. How LinkedIn Knows What Jobs You Are Interested In. UDACITY, 2014. Available online: https://blog.udacity.com/2014/05/how-linkedin-knows-what-jobs-you-are.html (accessed on 15 June 2019).
10. Claus, L. HR disruption—Time Already to Reinvent Talent Management. *BRQ Bus. Res. Q.* **2019**, in press, corrected proof. Available online: https://www.sciencedirect.com/science/article/pii/S2340943619302129 (accessed on 15 June 2019). [CrossRef]
11. Srivastava, R.; Palshikar, G.K.; Pawar, S. Analytics for Improving talent acquisition processes. In Proceedings of the International Conference on Advanced Data Analysis, Business Analytics and Intelligence, Ahmedabad, India, 11–12 April 2015. ICADABAI 2015.
12. Dutta, D.; Mishra, S.; Manimala, M.J. *Talent Acquisition Group (TAG) atHCL Technologies: Improving the Quality of Hire Through Focused Metrics*; Technical Report; IIMB-HBP: Bengaluru, India, 2015; Available online: http://research.iimb.ernet.in/handle/123456789/6698 (accessed on 15 June 2019).
13. Faliagka, E.; Tsakalidis, A.; Tzimas, G. An Integrated e-Recruitment System for Automated Personality Mining and Applicant Ranking. *Int. Res.* **2012**, *22*, 551–568. Available online: https://www.emeraldinsight.com/doi/abs/10.1108/10662241211271545 (accessed on 15 June 2019). [CrossRef]
14. Faliagka, E.; Ramantas, K.; Tsakalidis, A.; Tzimas, G. Application ofMachine Learning Algorithms to an online Recruitment System. In Proceedings of the ICIW 2012: The Seventh International Conference on Internet and Web Applications and Services, IARIA, Stuttgart, Germany, 27 May–1 June 2012; pp. 216–220.
15. Palshikar, G.K.; Srivastava, R.; Pawar, S.; Hingmire, S.; Jain, A.; Chourasia, S.; Shah, M. Analytics-Led Talent Acquisition for Improving Efficiency and Effectiveness. In *Advances in Analytics and Applications*; Springer: Singapore, 2019; pp. 141–160.
16. Mooney, R.J.; Bunescu, R. Mining Knowledge from Text Using Information Extraction. *ACM SIGKDD Explor. Newsl.* **2005**, *7*, 3–10. Available online: https://www.cs.utexas.edu/~{}ml/papers/text-kddexplore-05.pdf (accessed on 15 June 2019). [CrossRef]
17. Téllez-Valero, A.; Montes-y-Gómez, M.; Villaseñor-Pineda, L.A. Machine learning approach to information extraction. In *Computational Linguistics and Intelligent Text Processing. CICLing Lecture Notes in Computer Science*; Gelbukh, A., Ed.; Springer: Berlin/Heidelberg, Germany, 2005; pp. 539–547. Available online: https://link.springer.com/chapter/10.1007/978-3-540-30586-6_58 (accessed on 15 June 2019).
18. Tomassetti, F.; Rizzo, G.; Vetro, A.; Ardito, L.; Torchiano, M.; Morisio, M. Linked data approach for selection process automation in systematic reviews. In Proceedings of the 15th Annual Conference on Evaluation & Assessment in Software Engineering, Durham, UK, 11–12 April 2011; Institution of Engineering and Technology (IET): Sunnyvale, CA, USA, 2011; pp. 31–50.

19. Xu, Z.; Song, B.H. A machine learning application for human resource data mining problem. In Proceedings of the Advances in Knowledge Discovery and Data Mining, Singapore, 9–12 April 2006; Ng, W.K., Kitsuregawa, M., Li, J., Chang, K., Eds.; Lecture Notes in Computer Science. Springer: Berlin/Heidelberg, Germany, 2006; Volume 3918, pp. 847–856.
20. Xie, F.; Tang, Q. Human resource development by fuzzy neural networks. In Proceedings of the 4th International Conference on Wireless Communications, Networking and Mobile Computing, Dalian, China, 12–14 October 2008; IEEE: Piscataway, NJ, USA, 2008; Volumes 1–31.
21. Sivaram, N.; Ramar, K. Applicability of Clustering and Classification Algorithms for Recruitment Data Mining. *Inte. J. Comput. Appl.* **2010**, *4*, 23–28. Available online: https://pdfs.semanticscholar.org/22e3/1564b13413c537f246e7d59e9075df0db7f8.pdf (accessed on 15 June 2019). [CrossRef]
22. Aldarra, S.; Muñoz, E. A Linked Data-Based Decision Tree Classifier to Review Movies. In Proceedings of the Know@LOD 2015, 4th Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data co-located with 12th Extended Semantic Web Conference (ESWC 2015), Portoroz, Slovenia, 31 May 2015; Available online: http://ceur-ws.org/Vol-1365/paper10.pdf (accessed on 15 June 2019).
23. Mehenni, T.; Moussaoui, A. Data Mining from Multiple Heterogeneous Relational Databases Using Decision Tree Classification. *Pattern Recognit. Lett.* **2012**, *33*, 1768–1775. Available online: https://dl.acm.org/citation.cfm?id=2343166 (accessed on 15 June 2019). [CrossRef]
24. Sanchez-Marono, N.; Alonso-Betanzos, A.; Fontenla-Romero, O.; Polhill, J.G.; Craig, T. Empirically-Derived Behavioral Rules in Agent-Based Models Using Decision Trees Learned from Questionnaire Data. In *Agent-Based Modeling of Sustainable Behaviors. Understanding Complex Systems*; Alonso-Betanzos, A., Sánchez Maroño, N., Fontenla-Romero, O., Polhill, G.J., Craig, T., Bajo, J., Corchado, J.M., Eds.; Springer: Cham, Switzerland, 2017; pp. 53–76.
25. Bizer, C.; Heese, R.; Mochol, M.; Oldakowski, R.; Tolksdorf, R.; Eckstein, R. The Impact of Semantic Web Technologies on Job Recruitment. In Proceedings of the 7 Internationale Tagung Wirtschaftsinformatik, Bamberg, Germany, 23–25 February 2005.
26. Ristoski, P.; Petrovski, P.; Mika, P.; Paulheim, H. A Machine Learning Approach for Product Matching and Categorization. *Semant. Web* **2018**, 1–22, Preprint. Available online: http://www.semantic-web-journal.net/system/files/swj1470.pdf (accessed on 15 June 2019).
27. Ristoski, P.; Paulheim, H. Semantic Web in Data Mining and Knowledge Discovery: A Comprehensive Survey. *Web Semant. Sci. Serv. Agents World Wide Web* **2016**, *36*, 1–22. Available online: https://www.sciencedirect.com/science/article/pii/S1570826816000020 (accessed on 15 June 2019). [CrossRef]
28. Min, H.; Emam, A. Developing the Profiles of Truck Drivers for Their Successful Recruitment and Retention: A Data Mining Approach. *Int. J. Phys. Distrib. Logist. Manag.* **2003**, *33*, 149–162. Available online: https://www.emeraldinsight.com/doi/abs/10.1108/09600030310469153 (accessed on 15 June 2019). [CrossRef]
29. Gomez-Perez, A.; Fernández-López, M.; Corcho, O. *Ontological Engineering: With Examples from the Areas of Knowledge Management, E-Commerce and the Semantic Web*; Springer: Heidelberg, Germany, 2006.
30. Simperl, E. A Case Study in Building Semantic eRecruitment Applications. In *Semantic Web for Business: Cases and Applications*; Garcia, R., Ed.; IGI Global: Hershey, PA, USA, 2008.
31. Gregor, S.; Hevner, A.R. Positioning and Presenting Design Science Research for Maximum Impact. *MIS Q.* **2013**, *37*, 337–355. Available online: https://pdfs.semanticscholar.org/82a8/6371976aaf181a477745148eab07bb9ed143.pdf (accessed on 15 June 2019). [CrossRef]
32. The Source Code. Available online: https://github.com/catalinstrimbei/rdf-mining-hr (accessed on 15 June 2019).
33. Open Refine. Available online: http://openrefine.org/ (accessed on 15 June 2019).
34. Ontology Engineering Group, Human Resource Ontology. Available online: http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/ontologies/99-hrmontology/ (accessed on 15 June 2019).
35. Ericsson, K.A. The influence of experience and deliberate practice on the development of superior expert performance. In *The Cambridge Handbook of Expertise and Expert Performance*; Cambridge University Press: Cambridge, UK, 2006; Volume 38, pp. 685–705.
36. Langer, J.; Feeney, M.K.; Lee, S.E. Employee Fit and Job Satisfaction in Bureaucratic and Entrepreneurial Work Environments. *Rev. Public Pers. Adm.* **2019**, *39*, 135–155. Available online: https://journals.sagepub.com/doi/abs/10.1177/0734371X17693056 (accessed on 15 June 2019). [CrossRef]

37. Chambers, E.G.; Foulon, M.; Handfield-Jones, H.; Hankin, S.M.; Michaels, E.G. The War for Talent. *McKinsey Q.* **1998**, 44–57. Available online: http://www.executivesondemand.net/managementsourcing/images/stories/artigos_pdf/gestao/The_war_for_talent.pdf (accessed on 15 June 2019).
38. Bakotić, D. Relationship Between Job Satisfaction and Organisational Performance. *Econ. Res. Ekon. Istraž.* **2016**, *29*, 118–130. Available online: https://www.tandfonline.com/doi/full/10.1080/1331677X.2016.1163946 (accessed on 15 June 2019). [CrossRef]
39. Mahmood Ali, M.; Qaseem, M.; Rajamani, L.; Govardhan, A. Extracting Useful Rules Through Improved Decision Tree Induction Using Information Entropy. *Int. J. Inf. Sci. Tech.* **2013**, *3*. Available online: https://arxiv.org/ftp/arxiv/papers/1302/1302.2436.pdf (accessed on 15 June 2019). [CrossRef]
40. Melillo, P.; Orrico, A.; Chirico, F.; Pecchia, L.; Rossi, S.; Testa, F.; Simonelli, F. Identifying Fallers Among Ophthalmic Patients Using Classification Tree Methodology. *PLoS ONE* **2017**. Available online: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174083 (accessed on 15 June 2019). [CrossRef]
41. Davis, J.; Goadrich, M. The Relationship Between Precision-Recall and ROC Curves. In Proceedings of the 23 International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006.
42. Kalampokis, E.; Zeginis, D.; Tarabanis, K. On Modeling Linked Open Statistical Data. *J. Web Semant.* **2019**, *55*, 56–68. Available online: https://www.sciencedirect.com/science/article/pii/S1570826818300544 (accessed on 15 June 2019). [CrossRef]
43. Michel, F.; Zucker, C.F.; Corby, O.; Gandon, F. Enabling Automatic Discovery and Querying of Web APIs at Web Scale using Linked Data Standards. In Proceedings of the LDOW/LDDL Workshop of the 2019 World Wide Web Conference (WWW'19), San Francisco, CA, USA, 13–17 May 2019.
44. Modi, K.J.; Garg, S.; Chaudhary, S. An Integrated Framework for RESTful Web Services Using Linked Open Data. *Int. J. Grid High Perform. Comput.* **2019**, *11*, 24–49. Available online: https://www.igi-global.com/article/an-integrated-framework-for-restful-web-services-using-linked-open-data/224029 (accessed on 15 June 2019). [CrossRef]
45. Mochol, M.; Wache, H.; Nixon, L. Improving the Accuracy of Job Search with Semantic Techniques. In Proceedings of the 10th International Conference Business Information Systems, Poznan, Poland, 25–27 April 2007; Springer: Berlin/Heidelberg, Germany, 2007.
46. Kessler, R.; Torres-Moreno, J.; El-Beze, M. E-Gen: Automatic job offer processing system for human resources. In Proceedings of the Artificial Intelligence 6th Mexican International Conference on Advances in Artificial Intelligence (MICAI'07), Aguascalientes, Mexico, 4–10 November 2007; Rauch, J., Ras, Z., Berka, P., Elomas, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 985–995.