Case-Based Comparison of Career Trajectories

Kedma Duarte[1], Rosina O. Weber[2], Roberto C. S. Pacheco[1]

[1]Graduate Program in Knowledge Engineering and Management, Federal University of Santa Catarina, Brazil
kedmaduarte@gmail.com, pacheco@egc.ufsc.br
[2]College of Computing and Informatics, Drexel University, USA
rosina@drexel.edu

Abstract. Data generated across time may not be easily comparable in its original form thus potentially leading to results that may be perceived as unfair to some. We investigate quality assessment of scholarly researchers from their curricula vitae (CVs) for processes such as hiring, promotion, and grant funding. In previous work, we demonstrated that case-based reasoning (CBR) offers advantages as a transparent methodology to assess researcher quality. Its benefits include consistency, transparency, ability to adapt to specific purposes, and ability to provide explanation. The problem we now face is how to preprocess the data from the CVs to compare researchers whose scholarly production is achieved under different conditions, different points in time, and span different career trajectory lengths. We propose strategies to deal with these aspects of time during preprocessing of the data for case representation. We use 1,000 CVs from the Brazilian *Lattes* database to illustrate.

Keywords: case-based reasoning • time series • trajectory • career trajectory • curriculum vitae • normalization • recency

1 Introduction

There is a growing interest in relying on high quality profiling systems to conduct data studies to, as stated by Lane [1], "*make science more scientific*". Researcher quality assessment is a crucial task because characteristics of research metrics steer science and technology decisions, ultimately steering progress, economics, and our way of life [2].

Unfortunately, private organizations have started to explore this niche and are now steering our future by offering research metrics that rely on incomplete and flawed automatically crawled data [3]. In response to this present state, a group of researchers gathered at the 2014 International Conference on Science and Technology Indicators to produce the Leiden Manifesto [4]—a set of 10 principles for research quality metrics that includes attention to transparency, flexibility, and context, amongst others.

At the 2016 International Conference on Science and Technology Indicators, these authors proposed a CBR approach to manipulate profiling data for researcher quality assessment [5]. Our CBR method can be tailored to specific contextual purposes to meet some of the objective principles from the manifesto because of its consistency, transparency, ability to adapt to specific purposes, and ability to provide explanation.

In Proceedings of the ICCBR 2016 Workshops. Atlanta, Georgia, United States of America

In the proposed methodology, CBR is used to classify candidate researchers as either fit or unfit for a purpose. Purposes are characterized by features that reflect specific jobs or promotions. Each entails a series of references of quality such as publications in a journal or conference that are considered more relevant than others. The characterization of the purpose comes from the users who adopt the methodology to classify CVs of applicants. The use of CBR in this task assumes that assessing quality ultimately implies predicting future success.

In this paper, we describe the CBR implementation, and discuss three preprocessing steps that are required due to temporal aspects of the data. The first is a standard normalization step so that absolute volumes of scholarly production are replaced by relative values of productivity. This avoids the comparison of absolute numbers of production accomplished in years when conditions are different. The second aspect is recency. We analyze researchers' accomplishments to assess whether more recent production is or not more predictive of quality. The third refers to grouping the relative values of productivity depending on the lengths of career trajectories and recency. The CBR system, as it is implemented now, uses one aggregated data point for each attribute. Deciding how to group this data depends on directives of the users in terms of how they favor experience, productivity, or whether they want both to have the same emphasis.

This paper's intended contributions are to introduce the challenges stemming from using temporal data from CVs to assess researcher quality with CBR, and propose preliminary strategies to address them. We illustrate these challenges and strategies with data from 1,000 CVs from the period 2001 to 2014 from the Brazilian Lattes database [6]. The expected value of these strategies is to address these time-related challenges in a way that preserves transparency and enables an easy to understand substantiation.

In the next section, we provide the background for this work, including how we proposed to use CBR for assessing researcher quality. We also mention a few related works in time and CBR, and in time-series prediction. In Section 3, we describe the challenges and our proposed strategies. We lay out directions of future work in Section 4.

2 Background

In this section, we introduce some of the concepts used in this paper. We start with normalization, move to time-series approaches, and then discuss some aspects of dealing with career trajectories. In the final section, we describe the CBR approach that motivates this work.

Normalization is a method that may be used before a classification process, required to equalize ranges of the features from different scales, in order to obtain the same proportion between them, making features comparable [7]. Several techniques have been proposed to implement normalization (e.g., Min-Max Normalization, Linear Scaling to Unit Range, Median Normalization, and Z-Score Normalization), and many studies have investigated the relation between choosing the appropriated normalization technique and improving classification accuracy (e.g., [7][8]). These studies demonstrated the dependence of normalization methods in the performance of classification accuracy.

GenericPred [9] is a method for long-term time-series forecasting that addresses chaotic behaviors such as natural phenomena strongly dependent on initial conditions, which are many times unknown and consequently difficult to model and predict. The results of this approach demonstrated a significant gain in accuracy over traditional time series methods for both short and long-term predictions.

Time-series using bibliometrics data have been used to discover *distinguished* researchers [10]. Their approach is able differentiate researchers who have contributed a significant achievement amongst those publishing a few papers over a long period.

Time-series data of renal transplantation patients has been used in case-based binary classification [11]. The approach compares time series of creatinine courses using a distance measure based on linear regression.

Dynamic time warping (DTW) [12] is a distance measure to compare temporal se-quences based on dynamic programming. DTW is much more robust than measures based on the Euclidian distance [13] as it allows an elastic shifting of the time axis.

2.1 Career Trajectories

The terms career and trajectories are viewed as synonyms that describe the path from entering into the job market and its following steps [16]. Along the same lines, the career of a researcher has been described as a longitudinal account of an individual's productivity [17]. Our focus in this paper is on career trajectories from the perspective of the productivity of researchers along their careers [18][19].

The consideration of time when studying career trajectories is important for the reliability of indicators and rankings. Previous indicators or metrics that attempted to define a fixed interval of years have been highly criticized [20]. The purpose of normalizing time intervals and use annual productivity when assessing researcher quality is to make available the same transparent standards to all researchers who are assessed.

Our main problem is that this process must be transparent and able to substantiate its fairness. The assessment has to clearly consider and describe separately the biases that come from the description of purpose from the biases that originate in learning methods. The first issue we investigate is how to demonstrate whether an assessment can be fair when quality assessment is case-based, which requires comparison between researchers whose career trajectories span different intervals.

2.2 Purpose-Oriented Case-Based Researcher Quality Assessment

The purpose-oriented CBR approach classifies researchers as fit or unfit for the purpose of a target process (Fig. 1) such as hiring or promotion [5]. This method supports the Leiden Manifesto [4] to incorporate purpose in research metrics aligned with the con-cept that quality means fitness for purpose [14].

A purpose-oriented approach requires users to input the purpose as a set of standards or examples. For instance, for a target process to hire a researcher for the federal uni-versity of Rio de Janeiro to work with the Zika virus, publications in local conferences

where geographic issues are the focus may be considered of high importance when as-sessing quality of an applicant. Users can also indicate examples of fit and unfit re-searchers, which can then be used for weight learning.

The first parameter to be captured for a purpose p is the target interval of interest N, where $n \in$ N is the year in question within the interval of years N that are to be included in the data from candidates to be considered for a given purpose. Years y of importance are y_1 = Initial year, and y_n = Final year.

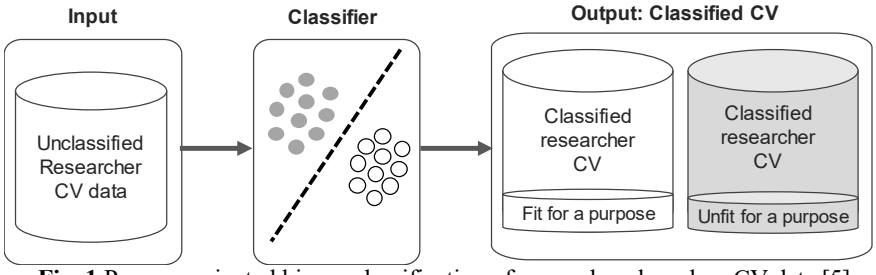

Fig. 1 Purpose-oriented binary classification of researchers based on CV data [5]

Cases $c_i \in$ C are researchers classified as fit or unfit for a purpose p. The set of cases C is defined as $C = (c_1, \ldots, c_m)$, where $i <= m$, and m represents the total number of cases under consideration. These cases comprise the case base CB.

Unclassified researchers are defined as $r_i \in$ R, where $R = (r_1, \ldots , r_m)$, with $i <= m$, as researchers are ultimately to receive a classification and become cases.

Cases are represented through its attributes by aggregated values within the inter-val N, at three stages. There is a finite set of attributes $a_{ij} \in$ A, $A = (a_{11}, \ldots, a_{mk})$, where $i <= m$, and k is number of attributes selected based on the raw CV data and the attributes of interest defined in the purpose, so $j <= k$. Attributes a_{ij} are used in the case-based approach to assess researcher quality.

Attributes $a_{ijn} \in$ A are values of the attributes from raw data before being subject to preprocessing for different years $n \in$ N. This is discussed in Section 3.1. Once normal-ized, a_{ijn} are converted into $\bar{a}_{ijn}$. Section 3.2 introduces recency, which produces a set o weights g_n for each year. Section 3.3 describes how to aggregate values $\bar{a}_{ijn}$ using weights g_n to compute aggregated attribute values a_{ij} introduced above.

These attributes are those typically found in profiling systems that have been shown to bear relevance to research related accomplishments [15]. Attributes a_{ij} include ac-complishments such as journal articles, published conference papers, and grant funding. Personal attributes such as age and gender that are not typical in quality assessments are not considered. These attributes are used to determine the fields of the CVs that are used and to capture the purpose. For example, in the arts field, one accomplishment may be artistic performances. When this is in the CVs, we need to capture how relevant different types of performances are considered for the purpose p (i.e., job).

Aggregated attributes a_{ij} are used in the CBR approach for similarity assessment, to compute a global similarity score *Global Sim*: U × CB → [0 , 1], where U is the universe of all objects CB from the case base:

$$Global\ Sim\ (r_i, c_i) = \sum_{j=1}^{k} w_j . sim(a_j, a'_j), 1 \leq j \leq k \qquad (1)$$

We define weights $W = (w_1, \ldots, w_m),\ W \in [0,1]$, for each purpose p.
The local similarity measure between attribute a_j, and a_j', used in the data in this article is defined by:

$$sim(a_j, a'_j) = \begin{cases} 1 - \left(\frac{|a - a'|}{d_{max}}\right), if |a - a'| < d_{max} \\ 0, else \end{cases} \quad (2)$$

where d_{max} is the maximum distance between a and a'.

This way the case-based classification of fit or unfit is not assessing similarity between time-series but between flat cases with weights in each attribute stemming from the characteristics of the purpose p.

3 Challenges and Proposed Strategies

We use a simple example to illustrate the challenges in preprocessing data to populate cases for case-based researcher quality assessment from CV data. Suppose we plan to use our case-based quality assessment approach to classify applicants as either fit or unfit for a given purpose. One of the parameters for a job is the target interval of interest N, which delimits the years that are considered relevant to include in the examination of candidates. For example, a job opening seeking social media experts would probably not include accomplishments from candidates that predate the existence of social media. For data in Table 1, the target interval of interest is five years. Each line in Table 1 refers to the volume of one type of accomplishment (e.g., journal articles in one field and reputation) produced by job applicants (i.e., researchers).

Table 1. Volume of one type of accomplishment produced by researchers

Researcher	Year 1		Year 2		Year 3		Year 4		Year 5	
	a_{ij1}	$\bar{a}_{ij1}$	a_{ij2}	$\bar{a}_{ij2}$	a_{ij3}	$\bar{a}_{ij3}$	a_{ij4}	$\bar{a}_{ij4}$	a_{ij5}	$\bar{a}_{ij5}$
$i = 1$	1	0.2	2	0.5	3	1	4	1	5	1
$i = 2$	5	1	4	1	3	1	2	0.5	1	0.2
$i = 3$	3	0.6	3	0.75	3	1	3	0.75	3	0.6
$i = 4$	0	0	0	0	3	1	3	0.75	3	0.6

3.1 Standard Normalization

Raw data attributes $a_{ijn} \in$ A are normalized using:

$$\bar{a}_{ijn} = \frac{a_{ijn}}{\max(a_{ijn})} \quad (3)$$

Suppose the data in Table 1 on the left columns designated by a_{ijn} reflect all the items produced by each applicant. The maximum number of accomplishments varies each year. Only in Year 1 and Year 5, a maximum of five accomplishments was produced. In Year 3 however the maximum produced was three. We contend that there are external factors that may have contributed to higher and lower levels of productivity. One common example is a reduction of participation in conferences in periods of economic depression. We therefore normalize these absolute values and convert them into productivity rates, using Equation 3.

The results of the normalization are laid out in right columns under $\bar{a}_{ijn}$. They show how one same absolute value (e.g., when $i = 3$) can represent the maximum productivity in Year 3 and 60% in Year 1. For example, if using absolute values, production of applicant in third row would be considered inferior to the applicant in the second row in Year 2 whereas relative values make them equal. This simple step is easy to describe to a broad audience and does not depend on characterization of the purpose.

3.2 Recency

The challenge with respect to recency stems from the notion that recent data may be perceived as more current and therefore more relevant in time series classification. This possible perception may lead to claims of injustice and therefore we need to establish a way to assess whether or not recent data is more influential. Note that in our proposed approach, the target process may dictate the importance of recent accomplishments. Assessing how influential recent data is would be required for implementations when the target process is neutral about recency.

Given our assumption that assessing quality implies predicting future success, it is consistent to interpret that data is influential or relevant when it is predictive. To do this, we take the target interval of interest N and set the last year aside as actual to provide outcome classes. The intuition is that if a given year's data has cases that correctly predict the actual year then this year's data are predictive and hence influential.

To demonstrate this proposed analysis, we start from a hypothetical purpose, namely, a job opening that seeks a researcher who is a successful collaborator. This hypothetical purpose was captured using rules that assigned more importance to publications and funded projects achieved in collaboration than to solo authored accomplishments. The data where we applied these rules to determine who was fit or unfit for a collaborative job was selected from the Brazilian Lattes database [6]. For this reason, some of the parameters used to create rules reflected that local culture. These data and weights were described in [5].

For the analysis we now describe, we use weights learned in [5]. The data we used in this analysis is new. We started from the entire Lattes database that retains around 4 million CVs. From these, we selected 212,000 CVs of researchers with completed doctoral degrees. In order to work with dense data, we kept only researchers who were continuously productive from the target interval of interest that we defined from 2001 to 2014, resulting in 50,000 CVs. We kept CVs from researchers with a growing absolute number of accomplishments to eliminate researchers with periods of inactivity. This resulted in 20,000 CVs. For the analysis we show in this paper, we used a randomly selected sample of 1,000 CVs.

For the target interval of interest from 2001 to 2014, we set aside 2014 as actual to provide outcome classes. Our goal is to assess how predictive the data from years 2001 to 2013 are. For each year, we use our case-based implementation with leave-one-out cross validation (LOOCV) [21] to predict whether each researcher would be classified as fit or unfit for the collaborative hypothetical purpose above described. We compute for each researcher whether the classification using each year is correct (i.e., true positive, true negative) or incorrect (i.e., false positive, false negative).

Table 2. Average accuracy (AA), accuracy of fit, and accuracy of unfit by year

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
AA	70.6	70.8	71.1	72.5	70.9	69.2	72.3	71.5	72.5	72.8	75.7	76.5	76.0
Fit	58.0	57.2	56.3	58.4	54.6	51.7	56.9	54.8	56.3	56.2	60.2	61.4	61.4
Unfit	72.8	73.8	75.6	77.4	77.6	76.7	78.7	80.3	81.0	82.5	84.8	85.1	83.6

Table 2 shows the average accuracy (AA) for all 1,000 researchers using their data from each year in the first row. The second and third rows present respectively accuracy of fit (i.e., ratio of true positives) and accuracy of unfit (i.e., ratio of true negatives).

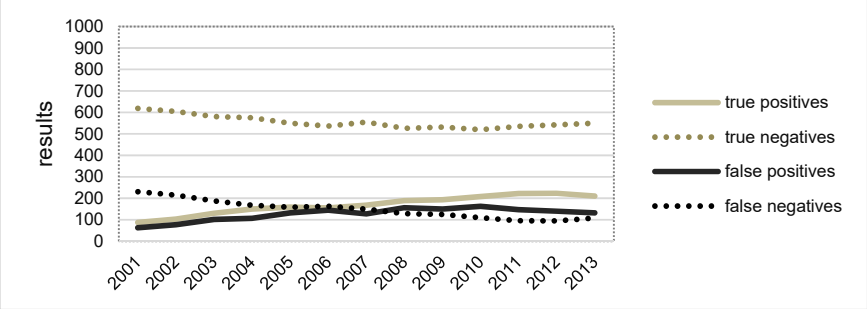

Fig. 2. True positives, true negatives, false positives, and false negatives in the target interval

These results in Table 2 are difficult to interpret because we do not know if the averages include the same or different researchers. To better understand these results, Fig. 2 plots true positives, true negatives, false positives, and false negatives. Positives are represented with continuous lines, and negatives are represented with dots. Light color is for true and dark for negative. A consistent trend would have true positives and true negatives in a direction opposite to false positives and false negatives. This would mean that accuracy increased because, for example, the true positives increased because of a reduction of false negatives. If accuracy increased in more recent years, then we would have to increase the relative relevance of these years when aggregating these values. The lines in Fig. 2 do not show consistent results. Our conclusion is thus that there is no consistent trend supporting the interpretation that recent data is more or less predictive than older data. Hence, $\forall\ \bar{a}_{ijn}$, N= {2001,..., 2013}, $g_n = 1$. Values for g_n are used in the next step when values are aggregated.

3.3 Career Trajectories

A researcher $r_i \in$ R has a career trajectory CT that reflects the researcher's years of activity. For aggregating researchers' production, we need:

$CT_i = y$

Max $(CT_i) = y_{max}$

Min $(CT_i) = y_{min}$

Table 3. Aggregated average when target process favors productivity

Researcher	Year 1	Year 2	Year 3	Year 4	Year 5	$\sum_{n=1}^{N} g_n \bar{a}_{ijn}$	$\frac{\sum_{n=1}^{N} \bar{a}_{ijn}}{N}$
$i = 1$	0.2	0.4	0.6	0.8	1	3	0.6
$i = 2$	1	0.8	0.6	0.4	0.2	3	0.6
$i = 3$	0.6	0.6	0.6	0.6	0.6	3	0.6
$i = 4$	0	0	0.6	0.6	0.6	1.8	0.6

The length of an applicant's career trajectory requires a decision that is part of the characterization of the purpose, specific to a given job opening. We need to ask whether the target recruiting process favors high productivity, experience, or both equally. Suppose we now examine a different attribute whose productivity values are in Table 3. The rightmost columns show the sum of production in all years and the average given by the sum divided by the years of activity. Note in our data zero production reflects a shorter career trajectory. The final average is the same for all applicants. This way of aggregating production through the years favors productivity and not experience. The example illustrates how a researcher with shorter trajectory is able to catch up with more experienced contenders.

Table 4. Aggregated average when target process favors experience

Researcher	Year 1	Year 2	Year 3	Year 4	Year 5	$\frac{\sum_{n=1}^{N} \bar{a}_{ijn}}{y_{min}}$
$i = 1$	0.2	0.4	0.6	0.8	1	1
$i = 2$	1	0.8	0.6	0.4	0.2	1
$i = 3$	0.6	0.6	0.6	0.6	0.6	1
$i = 4$	0	0	0.6	0.6	0.6	0.6

If the job were characterized as favouring experince, then we would not expect researcher $i = 4$ to have equivalent aggregated value. In this case, we propose to use the length of the shortest career of an applicant as the denominator to compute the average (Table 4). This would seem suitable, where the aggregated production for Candidate 4 is inferior to more experienced contenders.

Table 5. Aggregated average when target process favors equally productivity and experience

Researcher	Year 1	Year 2	Year 3	Year 4	Year 5	$\frac{\sum_{n=1}^{N} \bar{a}_{ijn}}{y_{max}}$
$i = 1$	0.2	0.4	0.6	0.8	1	0.6
$i = 2$	1	0.8	0.6	0.4	0.2	0.6
$i = 3$	0.6	0.6	0.6	0.6	0.6	0.6
$i = 4$	0	0	0.6	0.6	0.6	0.35

If the job is characterized as favoring productivity and experience equally, then we propose to use the length of the longest path as the denominator to compute the average. Note how the larger denominator used in Table 5 decreases the advantage of researcher $i = 4$ when compared to Table 3 and Table 4. The different denominators to aggregate these values will lead to greater results when the applicant is better suited to the characteristics of each job.

4 Concluding Remarks and Next Steps

This paper introduces time-related challenges faced when implementing CBR for researcher quality assessment. We propose a standard normalization to compare productivity instead of absolute volume of accomplishments, strategies to aggregate production across different career trajectories, and an analysis of predictiveness to address recency.

Given that there is no consensus on how many years should be used to assess researcher quality, we propose to use predictiveness of data within a target process context as a proxy to how influential it should be. We showed an illustrative example where data did not reveal variations in its level of predictiveness.

This work is very preliminary. The next step is to study different datasets to determine how to assess predictiveness and how to compute a measure of recency for when data reveals consistent trends.

The approach proposed in this paper aims to enhance the case-based researcher quality assessment proposed in [5] by adding weights within a target time interval when results of the recency assessment determine that more recent data is more or less predictive of the future and therefore should be considered more relevant.

Given that normalization strategies interfere with classification accuracy, we need to experiment with various purpose scenarios and normalization strategies to assess which have both high accuracy and acceptable substantiation. Along these lines, we will investigate DTW particularly when comparing career trajectories of different lengths.

This paper does not detail how the characterization of a purpose may be captured, which can be through examples, conditions, and a combination of these. We also limit the presentation to binary classification and do not discuss how to produce a ranking of the applicants. These are both topics for future work.

Acknowledgements

Authors thank the STELA Institute, particularly Rudger Taxweiler for his help collecting data. First author is supported by Brazilian's Goiás Research Foundation (FAPEG) and University of the State of Goiás (UEG) under agreement number 201310267000099. Authors also thank the suggestions from the reviewers.

References

1. Lane, J.: Let's make science metrics more scientific. Nature 464, 488–489 (2010)
2. Katz, J.S.: Scale-Independent Measures: Theory and Practice. In 17th International Conference on Science and Technology Indicators. Montreal, Canada, 1–19 (2012)
3. Van Noorden, R.: Metrics: A profusion of measures. Nature 465(7300), 864–866 (2010)
4. Hicks, D. Wouters, P., Waltman, L., De Rijcke, S., Rafols, I.: Bibliometrics: The Leiden Manifesto for research metrics. Nature, 520, 429–431 (2015)

5. Duarte, K., Weber, R., Pacheco, R.C.S.: Purpose-oriented metrics to assess re-searcher quality. In 21st International Conference on Science and Technology Indicators (STI2016): Peripheries, frontiers and beyond. València, Spain, 1312–1314 (2016)
6. Pacheco, R. C. S., Kern, V. M., Salm Jr, J. F., Packer, A. L., Murasaki, R., Amaral, L., Santos, L.D., Cabezas B., A. R.: Toward CERIF-ScienTI cooperation and interoperability. In: A.G.S. Asserson, E. J. Simons (Eds.) 8th International Conference on Current Research Information Systems. Leuven: Leuven University Press. 179–188 (2006).
7. Singh, B. K., Verma, K., Thoke, A. S.: Investigations on Impact of Feature Normalization Techniques on Classifier's Performance in Breast Tumor Classification. International Journal of Computer Applications, 116(19), 11–15 (2015)
8. Jayalakshmi, T., Santhakumaran, A.: Statistical Normalization and Back Propagation for Classification. International Journal of Computer Theory and Engineering, 3(1), 1–5 (2011)
9. Golestani, A., Gras, R.: Can we predict the unpredictable? Scientific reports, 4, 6834 (2014)
10. Kawamura, T., Yamashita, Y., Matsumura, K.: Research Activity Classification based on Time Series Bibliometrics. In 21st International Conference on Science and Technology Indicators (STI2016): Peripheries, frontiers and beyond. Valencia, Spain, 1456–1460 (2016)
11. Schlaefer, A., Schröter, K., Fritsche, L.: A case-based approach for the classification of medical time series. In International Symposium on Medical Data Analysis. Springer Berlin Heidelberg. 258–263 (2001)
12. Myers CS, Rabiner LR: A comparative study of several dynamic time-warping algorithms for connected word recognition. The Bell System Technical Journal 60(7):1389–1409 (1981)
13. Al-Naymat, G., Chawla, S., Taheri, J.: Sparse DTW: a novel approach to speed up dynamic time warping. In Proceedings of the Eighth Australasian Data Mining Conference. Australian Computer Society, Inc., 101, 117–127 (2009)
14. Juran, J. M, Godfrey, A. B.: Juran's Quality Handbook. New York: McGraw-Hill. 5th Edi-tion. (1999)
15. Duarte, K., Weber, R., Pacheco, R. C. S.: Conceptual data model for research collaborators. In CIKI: VI International Conference on Knowledge and Innovation (2016)
16. Valenduc G., Vendramin P., Pedaci M., Piersanti M.: Changing careers and trajectories. How individuals cope with organisational change and restructuring, WORKS re-port, HIVAK. U. Leuven, Leuven. (2009)
17. Dietz, J. S., Chompalov, I., Bozeman, B., Lane, E. O. N., Park, J.: Using the curriculum vita to study the career paths of scientists and engineers: An exploratory assessment. Scientometrics, 49(3), 419–442 (2000)
18. Lee, S., & Bozeman, B.: The impact of research collaboration on scientific productivity. Social studies of science. 35(5), 673–702 (2005)
19. Unger, D. D., Rumrill, P. D.: An Assessment of Publication Productivity in Career Development and Transition for Exceptional Individuals 1978–2012. Career Development and Transition for Exceptional Individuals, 36(1), 25-30 (2013)
20. Stuart, D.: Metrics for an increasingly complicated information ecosystem. Online Information Review, 39(6), 848–854 (2015)
21. Stone, M.: Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society. Series B (Methodological), 111–147 (1974)