Prediction of Employee Promotion Based on Personal Basic Features and Post Features

Yuxi Long
College of Systems Engineering
National University of Defense Technology
Changsha 410073, China
long_yuxi@163.com

Jiamin Liu
College of Systems Engineering
National University of Defense Technology
Changsha 410073, China
liujm_1020@163.com

Ming Fang
College of Systems Engineering
National University of Defense Technology
Changsha 410073, China
mingxfz@foxmail.com

Tao Wang
College of Systems Engineering
National University of Defense Technology
Changsha 410073, China
wangtao@nudt.edu.cn

Wei Jiang
Office of Reserve Officer Selection and Training
Stationed in Peking University and Tsinghua University
Beijing 100084, China
jwei0315@163.com

ABSTRACT
Promotion is the focus of human resource management research. Because there are few researches about the mining of promotion features in existing studies, this paper uses the data of a Chinese state-owned enterprise, constructs a number of features and applies machine learning methods to predict employee promotion. Firstly, we build personal basic features and post features based on five strategies. Secondly, the correlation analysis is conducted to preliminarily explore the associations between some features and promotion. Then, the model learning and testing are carried out. Experimental results show that the random forest model performs best, which verifies the validity of features. Finally, we calculate the Gini importance of each feature to further analyze its influence on staff promotion. It is found that post features have a higher impact on promotion compared with personal basic features. Among all the features, the working years, the number of different positions and the highest department level greatly affect employee promotion.

CCS Concepts
• Information systems→Data mining • Social and professional topics→Employment issues • Computing methodologies→ Supervised learning by classification.

Keywords
Employee promotion prediction; machine learning; personal basic features; post features

1. INTRODUCTION
As global competition intensifies, the competition among various enterprises becomes more and more fierce. Employees are the key part of enterprises and significantly affect enterprises development. Promotion is a concern for both enterprises and employees. On one hand, promotion is the approach that enterprises use to select outstanding talents and enhance their competitiveness. There is a strong correlation between employee promotion system and organizational performance [1]. On the other hand, it is also a way for staff members to realize self-worth and obtain development opportunities. Promotion opportunities have a greater effect on employee performance [2], the lack of which is likely to result in employees' voluntary turnover [3]. Therefore, an effective promotion management system is particularly vital for enterprises to attract, retain and use the talent. Moreover, it is of great significance to research promoted employees' characteristics and make predictions.

Promotion has always been a research hotspot in the field of human resource management. Some studies show that internal promotion in enterprises is affected by lots of factors, such as gender [4]-[6], age [7], [8], education background [9], [10], job experience [11], emotional intelligence [12] and communication patterns [13], [14]. In terms of research methods, the main approaches are qualitative analysis [15], quantitative analysis [16] and their combination. However, traditional studies are usually based on the data obtained from questionnaires and interviews, the results of which are probably to be influenced by subjective factors. In recent years, techniques of data mining or machine learning have been gradually applied to this field and have become increasingly prominent in the era of big data. For example, Yuan *et al.* [17] revealed the correlation between employees' social/work-related actions and their career development by building the action network (AN) and social network (SN), and a logistic regression model was adopted to predict promotion and resignation. Fan *et al*. [18] used the mixed neural networks and cluster analysis to predict the turnover of technology professionals. For solving the problem of human resources allocation, Xu and Song [19] utilized machine learning algorithms and proved their effectiveness in dynamic environment. A feature selection method based on neighborhood propagation clustering and SVM sensitivity analysis was put forward by Wang *et al*. [20], which was also successfully applied to human resource management.

ICDPA 2018, May 12–14, 2018, Guangdong, China

ACM ISBN 978-1-4503-6418-8/18/05…$15.00
DOI: https://doi.org/10.1145/3224207.3224210

Although some achievements have been made by applying big data analysis technology in human resource management, research on the mining of promotion features is relatively less and there is a need for further study. Thus, based on the data from a Chinese state-owned enterprise, we construct some promotion features and make forecasts with machine learning methods.

The remaining paper is organized as follows. Section 2 presents the description of the data set. Feature construction is introduced in detail in Section 3 and correlation analysis is presented in Section 4. Section 5 describes the model learning and testing. Feature importance analysis is shown in Section 6. Finally, conclusions and discussion are made in Section 7.

2. DATA DESCRIPTION

The experimental data set comes from the staff database of a Chinese state-owned enterprise. The database contains detailed information of 77218 employees. Each employee has at least one complete record, which includes personal basic information, position information, performance assessment information and other information. The enterprise makes new statistics on employees' all information every three months. If some information changes, a new record will be added for this employee. We extract personal basic information data and position information data from the database. Table 1 lists 13 attributes in the data set. Then, we delete missing values and outliers. Eventually, the data set involves 71132 employees. In addition, some categorical attributes that were not encoded when stored in the database are recoded with numeric values. For some ordinal attributes, such as the department level and the position level, smaller numbers represent higher grades.

Table 1. Attributes of personal basic information data and position information data

Data	Attribute	Example
Personal basic information data	Gender	Male
	Birth date	1970/1/1
	Degree of education	Master
	Hometown (province)	Beijing
	Nationality	Han
Position information data	Work department	Finance Department
	Department level	1
	Position	Director
	Position level	3
	Position type	Managerial post
	Personnel nature	Regular worker
	Start time of current position	2010/1/1
	Record time	2010/12/30

3. FEATURE CONSTRUCTION

In this paper, we combine each employee's multiple records into a new set. The set of employee i is represented as $A_i = \{A_{i1}, A_{i2}, \cdots, A_{i13}\}$, where $A_{ij} = (a_{ij1}, a_{ij2}, \cdots, a_{ijn}), j=1,2,\ldots,13$ is the sequence that contains all record values of the j-th original attribute for employee i, and a_{ijn} denotes the n-th record value. The number of records n for each employee may be different and values in sequence A_{ij} are sorted by record time. Then, two types of features, i.e. the personal basic features and the post features, are constructed based on following five strategies.

(1) Strategy 1: choose the unique value.

For some attributes of the personal basic information data, like gender and nationality, values in their corresponding sequences are the same. Hence, we take this kind of attribute directly as a feature and use the unique value as the feature's value.

(2) Strategy 2: choose the mode.

In the sequence, some values may appear more than once, and the mode generally can represent the main value. This strategy refers to finding the mode of the attribute sequence A_{ij} and using it to denote the value of new feature. Main position, main department and other similar features can be built by this strategy.

(3) Strategy 3: choose the maximum or minimum value.

This strategy is suitable for the attribute whose values are ordinal and the optimal value is usually the maximum or minimum value in the sequence. For example, suppose $A_{i1} = (a_{i11}, a_{i12}, \cdots, a_{i1n})$ denotes the sequence of the department level for employee i, and F_8 denotes the highest department level, then the value of feature F_8 for employee i is defined as follows:

$$F_{i8} = min\ a_{i1k},\quad k=1,2,\ldots,n \tag{1}$$

(4) Strategy 4: count the number of different values.

For the working department, position and position type, counting the number of different values in their corresponding sequences is significant, because this kind of value may indicate the richness of employees' work experience and some studies have revealed that work experience has an impact on promotion. For instance, the number of different positions that an employee has engaged in may reflect his ability to work. The more positions he has worked in, the more experience he may accumulate, and then his competence will be stronger [21].

(5) Strategy 5: calculate the difference between two dates.

We build two features about time according to this strategy, respectively are the age and the working years. The age of each staff member refers to the age at which he was hired into the company. We assume that $A_{i2} = (a_{i21}, a_{i22}, \cdots, a_{i2n})$ is the sequence of start time of current position, $A_{i3} = (a_{i31}, a_{i32}, \cdots, a_{i3n})$ is the sequence of birth date and a_{ijn} (j=2,3) is the year of each date. The age is denoted as F_2, then the age of employee i can be calculated as follows:

$$F_{i2} = a_{i21} - a_{i31} \tag{2}$$

Similarly, let F_{12} represent the working years at this company, the value of feature F_{12} for employee i is defined as:

$$F_{i12} = a_{i2n} - a_{i21} \tag{3}$$

Table 2 shows the features we obtain and their corresponding construction strategies. Combining these features, we get the feature vector $\vec{X} = (F_1, F_2, \cdots, F_{15})$. Besides, the promotion label for each employee is built based on the position level. We assume that $A_{i4} = (a_{i41}, a_{i42}, \cdots, a_{i4n})$ denotes the sequence of position

level and calculate the increment of position level for employee i according to the following formula:

$$add\_pl_i = max\ a_{i4p} - min\ a_{i4q}, \quad p, q = 1, 2, \ldots, n \qquad (4)$$

Then, the employee whose position level increases obviously is classified as the promoted employee, while the employee who does not have a significant increase in position level is classified as the non-promoted employee. Through statistics, it is discovered that the maximum increment of position level among all employees is 4. Hence, the promotion label of employee i is defined as:

$$y_i = \begin{cases} 0 & \text{if } add\_pl_i < 2 \\ 1 & \text{if } add\_pl_i \geq 2 \end{cases} \qquad (5)$$

After building labels for all samples, we find that the ratio of promoted employees to non-promoted employees is close to 1: 1.

4. CORRELATION ANALYSIS

In order to initially explore the influence of personal basic features and post features on promotion, we select some typical features from the feature vector, and analyze the correlations between these features and promotion. In this paper, we define the promotion rate as the measurable indicator. Taking gender as an example, male employees' promotion rate equals to the ratio of the number of promoted male staffs to the total number of male staffs, and the promotion rate of female employees can be obtained similarly. Figure 1 and Figure 2 respectively show the promotion rate distributions of four personal basic features and four post features.

Table 2. Features and corresponding construction strategies

Feature type	Symbol	Feature	Strategy
Personal basic features	F_1	Gender	1
	F_2	Age	5
	F_3	Degree of education	3
	F_4	Hometown	1
	F_5	Nationality	1
Post features	F_6	Main department	2
	F_7	Main department level	2
	F_8	Highest department level	3
	F_9	Main position	2
	F_{10}	Main position type	2
	F_{11}	Main personnel nature	2
	F_{12}	Working years	5
	F_{13}	Number of different positions	4
	F_{14}	Number of different position types	4
	F_{15}	Number of different departments	4

As shown in Figure 1, the promotion rate of male employees in this enterprise is significantly higher than that of female employees. Nevertheless, promotion rates of Han employees and minority employees are not much different. As for the age, staff members aged 20-29 are more likely to be promoted than those aged 40-50. Employees with master's degrees have the highest promotion rate compared with employees who hold bachelor's or doctoral degrees. Based on above results, we can guess that gender might be more relevant to promotion. In addition, age and education may also be related to promotion, while nationality may have less impact on promotion.

By analyzing the results of Figure 2, we find there is a big difference between the promotion rate of employees who have worked in this enterprise for more than 10 years and that of employees who have worked for less than 5 years. As regards the number of different positions and the number of different departments, staffs who have worked in more than four positions or departments have a higher promotion rate. Besides, employees whose highest department level is 1 have a greater chance of getting promotion. Among the above four features, the differences in promotion rates of employees with disparate numbers of departments are relatively smaller. Therefore, we can think the working years, the number of different positions, the number of different departments and the highest department level may all have correlations with promotion, but the number of different departments has less influence on promotion than the other three features.

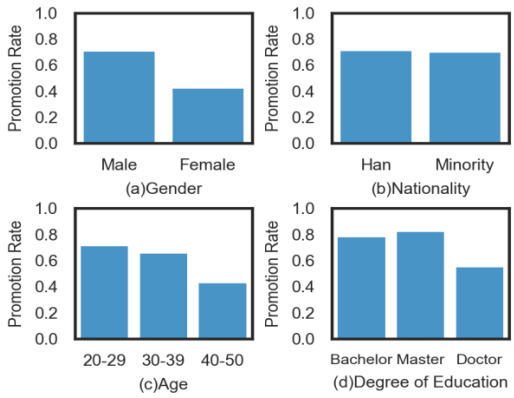

Figure 1. Promotion rate distributions of four personal basic features.

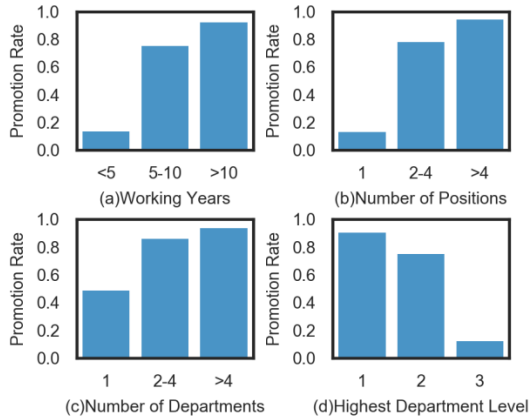

Figure 2. Promotion rate distributions of four post features.

5. MODEL LEARNING AND TESTING

To make features meet the requirements of machine learning, we adopt the min-max normalization to deal with numerical features

and utilize the One-Hot encoding to transform discrete features. The feature vector is extended to 100 dimensions after encoding and we use $\overrightarrow{X'}$ to denote the extended feature vector. Moreover, the data set is divided into the training set and the test set according to the ratio of 8: 2. For verifying the validity of features, we carry out model selection, parameter adjustment and model testing following the general process of machine learning.

5.1 Model Selection

Common classification models mainly include k-nearest neighbor (KNN), logistic regression (LR), support vector classifier (SVC), decision tree (DT), random forest (RF) and Adaboost. We take these six models as alternative models and adopt the k-fold cross-validation method to realize the initial selection of models. At the same time, we also compare the training effects of different feature subsets or feature vectors on each model, and the results are shown in Figure 3. In this figure, the vertical axis represents the cross-validation accuracy score. The feature subset *X1* is the set of personal basic features, *X2* is the set of post features, and *X3* is the feature vector $\overrightarrow{X'}$. It can be seen that the training effect of *X3* is better than that of *X2* and *X1* for each model, and the training performance of *X1* is relatively weak. These results indicate that the impact of post features on promotion is greater than that of personal basic features, and using these two types of features together for model training performs best. Overall, the prediction effects of LR, SVC, RF and Adaboost are better for the problem and the data used in our research.

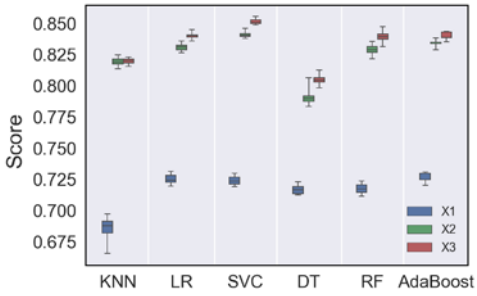

Figure 3. Cross-validation accuracy scores of models.

5.2 Parameter Adjustment

Since models' hyper-parameters also affect their performance, we adjust the parameters of selected four models through grid search and cross-validation. The core idea of this method is to set several combinations of parameters in advance and perform cross-validation for each set of parameters, so as to find the optimal parameter combination with the highest average score of cross-validation. Table 3 lists hyper-parameters values and corresponding cross-validation average scores of four models.

Table 3. Hyper-parameters values and cross-validation average scores of models

Model	Hyper-parameter	Value	Score
LR	C	100	0.892
	solver	lbfgs	
SVC	C	1000	0.894
	gamma	0.01	
RF	n_estimators	400	0.902
	max_depth	25	
Adaboost	learning_rate	1.0	0.892
	n_estimators	250	

5.3 Model Testing

In this part, the training set corresponding to feature vector $\overrightarrow{X'}$ is adopted to train these four classification models, and the test set is used to examine the validity of models. Since the problem studied in this paper is a two-classification problem, we measure the performance of each model by the receiver operating characteristic (ROC) curve and calculate the area under the curve (AUC). As shown in Figure 4, the prediction effect of RF is the best and its AUC value reaches 0.96.

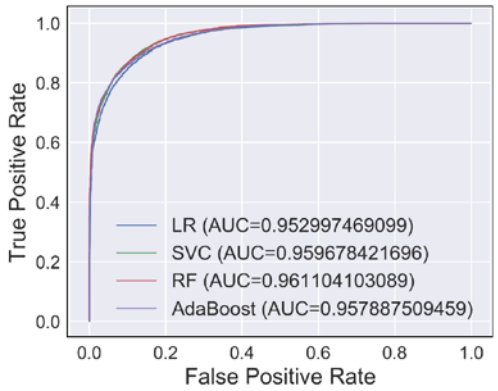

Figure 4. ROC and AUC of models.

6. FEATURE IMPORTANCE ANALYSIS

To further analyze the importance of features, the Gini importance of each feature in $\overrightarrow{X'}$ is calculated using the RF model. The higher Gini importance value of the feature, the greater its impact on the prediction. By accumulating the Gini importance of features encoded from same original categorical features, the ranking result of 15 original features' Gini importance can be obtained, as shown in Figure 5.

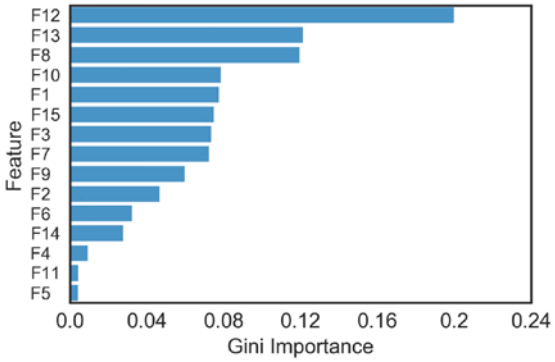

Figure 5. Gini importance of features.

The top three features in the figure are the working years (F_{12}), the number of different positions (F_{13}) and the highest department level (F_8). Among all the personal basic features, the importance of gender is relatively larger. As a whole, the importance of personal basic features is smaller than that of post features. The above analysis results show that: (1) The working years has a

relatively higher impact on employee promotion, followed by the number of different positions and the highest department level. (2) The effect of gender on promotion is greater than that of other personal basic features. (3) Post features have a bigger influence on promotion compared with personal basic features.

7. CONCLUSIONS AND DISCUSSION

In this paper, we extract staff members' personal basic information data and position information data from the employee database of a Chinese state-owned enterprise. Two types of features are constructed based on five strategies, and then we verify the effectiveness of features in forecasting employee promotion by correlation analysis, model learning and testing as well as feature importance analysis.

According to the results of correlation analysis, we speculate that some features might be associated with promotion, such as the gender, the working years and the number of different positions. Through model learning and testing, it is discovered that the prediction effect of random forest model is relatively better because its AUC value reaches 0.96, which proves the validity of features. The results of feature importance analysis further indicate that the impact of post features on promotion is greater than that of personal basic features, and the working years, the number of different positions and the highest department level have relatively larger influence on promotion. In addition, conclusions reached by correlation analysis and feature importance analysis are basically consistent. Among all the features, the importance of working years is the highest, and there are also big differences between the promotion rates of employees with different working years in the correlation analysis. It makes sense that employees with longer working years are more likely to be promoted to higher-level positions, because they may have made more contributions to the enterprise. The number of different positions has the second greatest importance, and it is found that employees who have experienced more than four positions have a higher promotion rate. This result is also reasonable, because such employees generally have richer experience and stronger ability. As for the highest department level, employees who have worked in higher-level departments are possible to have better position levels eventually, which may be related to the promotion management regulations of the enterprise. Furthermore, the ranking result that the Gini importance of gender is higher than that of other personal basic features also proves there is a certain association between gender and staff career development. This can be explained by the phenomenon that some enterprises are more willing to choose male employees for specific positions, although the capacity of some female employees may be stronger.

Many other factors, such as the family background, awards and punishments, may also affect the promotion of employees in companies. Therefore, in the next step of the research, we will continue to explore more features that have strong correlations with the promotion problem. In addition, we will try to further study more complex issues related to promotion. For example, we can attempt to predict the promotion speed or study whether a promoted employee is qualified for the higher-level position, and then put forward more relevant management suggestions for enterprises.

8. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No. 61773120, 71701206, 61473301, 61503402 and 71701205).

9. REFERENCES

[1] Lai, H. H. 2012. Study on influence of employee promotion system on organizational performance. *International Journal of organizational Innovation*. 5, 1 (July. 2012), 231-251.

[2] Nguyen, P. D., Dang, C. X. and Nguyen, L. D. 2015. Would Better Earning, Work Environment, and Promotion Opportunities Increase Employee Performance? An Investigation in State and Other Sectors in Vietnam. *Public Organization Review.* 15, 4 (Dec. 2015), 565-579. DOI= https://doi.org/10.1007/s11115-014-0289-4.

[3] Tanton, S. N. 2007. *Talent management in the role of employee retention*. Doctoral dissertation, University of South Africa.

[4] Blau, F. D. and DeVaro, J. 2007. New evidence on gender differences in promotion rates: An empirical analysis of a sample of new hires. *Industrial Relations: A Journal of Economy and Society.* 46, 3 (July. 2007), 511-550. DOI= https://doi.org/10.1111/j.1468-232x.2007.00479.x.

[5] Song, Y. 2007. Does Gender Make a Difference?-Career Mobility in Urban China. *China Economic Quarterly*. 6, 2 (Jan. 2007), 629-654. DOI= https://doi.org/10.13821/j.cnki.ceq.2007.02.007.

[6] Roth, P. L., Purvis, K. L., and Bobko, P. 2012. A meta-analysis of gender group differences for measures of job performance in field studies. *Journal of Management.* 38, 2 (Mar. 2012), 719-739. DOI= https://doi.org/10.1177/0149206310374774.

[7] Machado, C. S. and Portela, M. 2013. Age and opportunities for promotion. *IZA Discussion Paper No.7784.*

[8] Adams, S. J. 2002. Passed over for promotion because of age: an empirical analysis of the consequences. *Journal of Labor Research*. 23, 3 (Sep. 2002), 447-461. DOI= https://doi.org/10.1007/s12122-002-1046-y.

[9] Spilerman, S., and Lunde, T. 1991. Features of educational attainment and job promotion prospects. *American Journal of Sociology.* 97, 3 (Nov. 1991), 689-720. DOI= https://doi.org/10.1086/229817.

[10] Bognanno, M. L. and Melero, E. 2016. Promotion signals, experience, and education. *Journal of Economics & Management Strategy.* 25, 1 (Spring 2016), 111-132. DOI= https://doi.org/10.1111/jems.12132.

[11] De Pater, I. E., Van Vianen, A. E., Bechtoldt, M. N., and KLEHE, U. C. 2009. Employees' challenging job experiences and supervisors' evaluations of promotability. *Personnel Psychology.* 62, 2 (Summer 2009), 297-325. DOI= https://doi.org/10.1111/j.1744-6570.2009.01139.x.

[12] Kulkarni, P. M., Janakiram, B., and Kumar, D. N. S. 2009. Emotional intelligence and employee performance as an indicator for promotion, a study of automobile industry in the city of belgaum, karnataka, india. *International Journal of Business and Management.* 4, 4 (Apr. 2009), DOI= https://doi.org/10.5539/ijbm.v4n4p161.

[13] Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N. and Malone, T. W. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science*. 330, 6004 (Oct. 2010), 686-688. DOI= https://doi.org/10.1126/science.1193147.

[14] Pentland, A. 2012. The new science of building great teams. *Harvard Business Review.* 90, 4 (Apr. 2012), 60-69.

[15] Ridder, H. G. and Hoon, C. 2009. Introduction to the special issue: qualitative methods in research on human resource management. *German Journa of Human Resource Management: Zeitschrift für Personalforschung.* 23, 2 (May. 2009), 93-106. DOI= https://doi.org/10.1177/239700220902300201.

[16] Sanders, K., Cogin, J. A., and Bainbridge, H. T. J. 2014. *Research* methods *for human resource management.* Routledge.

[17] Yuan, J., Zhang, Q. M., Gao, J., Zhang, L. Y., Wan, X. S., Yu, X. J. and Zhou, T. 2016. Promotion and resignation in employee networks. *Physica A: Statistical Mechanics and its Applications.* 444 (Feb. 2016), 442-447. DOI= http://dx.doi.org/10.1016/j.physa.2015.10.039.

[18] Fan, C. Y., Fan, P. S., Chan, T. Y. and Chang, S. H. 2012. Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. *Expert Systems with Applications*. 39, 10 (Aug. 2012), 8844-8851. DOI= https://doi.org/10.1016/j.eswa.2012.02.005.

[19] Xu, Z. and Song, B.H. 2006. A machine learning application for human resource data mining problem. *Advances in Knowledge Discovery and Data Mining.* 3918 (2006), 847-856. DOI= https://doi.org/10.1007/11731139_99.

[20] Wang Q. W., Li B. Y. and Hu J. L. 2009. Feature selection for human resource selection based on affinity propagation and SVM sensitivity analysis. In *Proceedings of 2009 World Congress on Nature & Biologically Inspired Computing* (Coimbatore, India, December 9 - 11, 2009). NaBIC '09. IEEE Computer Society Press, 31-36. DOI= https://doi.org/10.1109/NABIC.2009.5393596.

[21] Dragoni, L., Oh, I. S., Vankatwyk, P., and Tesluk, P. E. 2011. Developing executive leaders: The relative contribution of cognitive ability, personality, and the accumulation of work experience in predicting strategic thinking competency. *Personnel psychology*. 64, 4 (Nov. 2011), 829-864. DOI= https://doi.org/10.1111/j.1744-6570.2011.01229.x.