AIC 2018: GLOBAL DEVELOPMENT IN HUMANITIES, EDUCATION AND CIVILIZATION

Using Data Mining to Predict Academician Publication Output

Mohd Zakree bin Ahmad Nazri*

Centre for Profesional and Leadership Development, Universiti Kebangsaan Malaysia
Malaysia

Rohayu Abdul Ghani

Centre for Profesional and Leadership Development, Universiti Kebangsaan Malaysia
Malaysia

Mas Ayu Che Razali

Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia,
Malaysia

**Corrospoding author's Email: zakree@ukm.edu.my*

*Peer-review under responsibility of 4th Asia International Conference 2018 editorial board
(<http://www.utm.my/asia/our-team/>)*

*© 2018 Published by Readers Insight Publisher,
lat 306 Savoy Residencia, Block 3 F11/1, 44000 Islamabad. Pakistan,
info@readersinsight.net*

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Research Highlights

This research focuses on predicting academician performance in terms of publication productivity and investigate the factors that affect academicians' achievement. Previous studies have shown that there are many important variables when analysing academicians' productivity at the individual level. This study investigates how scientific publication rate by individual is influenced by factors such as gender, age, number of research grant and academic position of the researchers using decision tree. Having a decision rules, university leaders can understand upcoming trends with respect to leadership requirements and academicians needs. It is also helping university managements understand challenges and therefore can deploy the right strategies for human resource management interventions. The discovered knowledge among the attributes obtained from mining the data can be used to predict the university productivity output. The study, involving almost 3000 university lecturers, shows a number of interesting patterns that can be used for predicting publication output.

Research Objectives

The main question for this research is to answer question - Is it possible to predict academician performance in publication? This study aims to fulfil the human resource (HR) requirement for a predictive model that can be used to 'see the future' of the university's output. The model can be used and applied in many ways such as a guidance for HR to produce specific recruitment and HR management strategies. The second objective of this study is to investigate factors associated with scholarly publication productivity among academic staff by using data mining to find unseen influence of attributes (factors) associated with scholarly publication productivity among academic staff.

Methodology

The predictive modelling in this paper uses the Cross Industry Standard Process for Data Mining (CRISP-DM) model. CRISP-DM is chosen as the methodology because it is the most widely-used analytic process standard (Brown, 2015). Shearer (2000) structured the CRISP-DM process model into six major phases: i.) Business Understanding; ii) Data

Understanding; iii. Data Preparation; iv. Modeling and v. Evaluation. However, in this research, the model is still being studied and therefore will not be deployed until we have rebuild the model with other state-of-the-art methods. The extract, transform and load (ETL) processes were involved in extracting data from various databases. The data set used in this study contains information from Student Information Systems, Human Resource Management Systems, Publication Management Systems, Leave Management Systems and Research & Innovation Management Systems. Before the loading take place, the data steward followed the process of Privacy Preserving Data Mining (PPDM) by removing all sensitive private information for each record. The dataset are collected from the university's information systems in collaboration with the CIT and Human Resource Department (HRD). These data were extracted from various databases and stored in a data warehouse before it was pre-processed.

Results

The output of modeling or data mining using decision tree is classifier. The results of the four decision tree techniques are as follow: The accuracy of predictive model using Decision Tree is 70.3%, PART's accuracy is 75%, J-48 obtained 75.3% and C4.5 received 70.2%. All these percerntahe represents the evaluation of the performance of the classifiers. The best decision trees performance obtained an accuracy of approximately 75%. However, it did not present any significant performance gain when compared to each other. Using the classifier produced by Decision Tree, 70.31% of all predictions done by this model are correct. However the outcome is most likely Need Improvement (NI). In fact, the confidence for this decision is only 70.24%. Compared to other methods, we found that PART is a better technique. The classifier generated by PART is with high-confident that the correct prediction is for the class of Meet Expectation (ME). The confidence for this decision is high with 96.15%.

Findings

This research focus on the development of predictive models for predicting and evaluating academic staff performance in the function of performance management. The biggest support for the rules obtained in this study is coming from position or rank in the academic. Hemmings and Kay (2010) finding support this pattern. They state that one of the important

factor associated with qualifications and frequently linked to research output is academic rank (Hemmings and Kay, 2010). 74.66% of all predictions done by classifier generated by PART are correct. When the classifier model predicts ME, it covers 72.39% of those cases and it is correct with 65.26% of all predictions for class ME. Based on the PART model, the research found out that several factors had a great effect on publication output. Below are of the most effective factors for predicting ME:

1. Academic Rank (e.g. Professorship).
2. Number of Research Grant as Member (Active + Completed Project),
3. Number of Masters Student (Supervised)
4. Number of Research Grant as the Head of Project (Active + Completed Project).
5. Number of Dependants (e.g, number of children).
6. Number of Conference Attendance as Presenter.
7. Number of Leaves at the Beginning of the Year (Entitlement per year + Accumulation).

Based on the analysis of the results, the study concluded that the researcher's experience as a speaker at conference, number of research grant both as member and head of project was the variable that had the most influence on research output besides the researcher's academic position. Since the study was conducted only with records of Professor and Associate Professor, the research has not conclude yet and will move to the next phase of this study to generalize it for the larger sample which consist of all academic staffs from this institution. Further research can therefore be performed for all academic staffs of the university or at other public universities in the country to consider the differences in the working environments of those institutes. Finally, investigations into the other factors associated with a scholarly publication productivity need to be carried out to further determine the output efficiency. Analysis on predictive model of academic staff performance is an important process for both human resource and the Deanery of each faculty in a university, since it can help a better understanding of what will happen and support decision maker to intervene with corrective measures and ultimately improve it.

Acknowledgement

This research was supported by Universiti Kebangsaan Malaysia under the KRA 4 Research Grant (KRA-2018-014 & KRA-2018-015). We thank our colleagues from the Human Resource Departments and UKM Centre for Information Technology (PTM) who provided

insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper.

References

- Brown, M.S., 2015. What IT Needs To Know About The Data Mining Process [WWW Document]. Forbes.
- Hemmings, B., Kay, R., 2010. University lecturer publication output: Qualifications, time and confidence count. *J. High. Educ. Policy Manag.* 32, 185–197.
- Shearer, C., 2000. The CRISP-DM model: the new blueprint for data mining,. *J. data Warehous.* 5, 13–22.