Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 160 (2019) 538–542

Procedia Computer Science

www.elsevier.com/locate/procedia

The 6th International Symposium on Emerging Information, Communication and Networks (EICN 2019)
November 4-7, 2019, Coimbra, Portugal

Big Data Aplication for Selecting Theses Topics

Jesus Silva[a]*, Lissette Hernandez[b], Tito Crissien[c], Omar Bonerge Pineda Lezama[d], Jenny Romero[e]

[a]*Universidad Peruana de Ciencias Aplicadas, Lima 07001, Peru*
[b,c,e] *Universidad de la Costa, Cl. 58 # 55 – 66, Barranquilla 080001, Colombia*
[d]*Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula 21101, Honduras*

Abstract

The chairs of thesis research in Computer Sciences at the University of Mumbai, in India, have different tools that provide academic support to those who begin the course. Students present difficulty in selecting the topic to study and, in some cases, this is a reason to neglect or delay in the career. Having a registry of thesis of the graduates proved to be useful for obtaining patterns on the relationship between the research area addressed in this thesis and the characteristics that define the researcher. For this study, the knowledge discovery process in Database was used. Finally, the results of the use of algorithms J4.8 in WEKA, ID3 tags in RapidMiner and CART IN Knime and the various models generated, with the data collected were presented.

Keywords: Decision trees, WEKA, RapidMiner, Knime, areas of research.

1. Introduction

Computer Science Engineering from the University of Mumbai, in India, consists of a course named "Degree Thesis - Integrating Final Project " [1]. The Bachelor´s degree in Systems has two quarterly subjects, being the first of them called "Diploma Work" and the second one called "Thesis" [2]. The thesis chairs of both careers do not have

* Corresponding author. Tel.: +51-975717503.
E-mail address: jesussilvaUPC@gmail.com

10.1016/j.procs.2019.11.051

a formal instrument to allows teachers identify the knowledge of students, their work experience, their times of dedication to the academy, their personal characteristics, the research lines, preferred development and innovation, residence times for the development of their thesis, among others. At present, the thesis chairs have a spreadsheet that contains information about student data (last name, first name, registration and career, the date of the thesis defense, teacher-tutor or teacher-director of research, title of research, research line, abstract, objectives, and future studies). This file allowed the chairs make a systematic registration of 452 theses from the year 2015 to March 2018 [3].

Over the last few years, teachers of the chairs of the two careers noted that the biggest drawback that the students manifest at the beginning of the course is the definition of the topic, causing a delay in the completion of their studies and, in some cases, the drop-out of the career. In this context, the following research questions raised for the study. Are data collected from graduates enough to identify patterns in the selection of thesis topics? Can the selected thesis area be correlated with features such as a work area, available time for developing the thesis, age at the beginning of the thesis, among others?

The answer to the first question caused the construction of a data collection instrument called THESIS STUDENTS-UM for completing the information of the chair. The quality of the instrument was tested, and the sample was formed by 117 graduates, including 49 students of Systems and 68 of Informatics Engineering [4]. For the second question, the process named Knowledge Discovery in Databases, (KDD) was applied since "the process is not trivial to identify valid, new, potentially useful and, understandable patterns from the data" [5]. This process involves several stages, ranging from obtaining the data until the implementation of the acquired knowledge in decision-making processes. Between these stages, Data Mining (DM) is the core of the KDD process [6].

The problems addressed in this study are part of the general outline of the KDD applying sorting tasks to obtain a descriptive model included in the Data Mining stage. In this stage, learning methods are applied for obtaining models and patterns. Learning always will be understood as supervised, where cases belonging to the data set have a priori class or category with the aim of finding patterns or trends in the cases belonging to the same class.

2. Method

For the discovery of the information, the proposed phases in [7]. were applied [8], which are explained in the Results section.

The spreadsheet of the chair was used allowing a systematic record of the defended theses and the built data collection instrument [5]. At this stage of the study, a sample was composed of 3,850graduates of Systems and 6,895 graduates of Informatics Engineering, for a total sample of 10,745.

Table 1 presents the attributes to be used in the DM phase with the associated values after the transformation and data cleaning.

Table 1. Attributes and values to be used in the Data Mining phase

Attributes	Values
Thesis Area	Agents and Intelligent Systems/ Software Engineering/ Database and Data Mining/ Innovation in Software Systems/ Architecture, Networks and Operating Systems/ Security/ Technology and Education/ Processing of Signals and Systems in Real Time
Career	Computer Engineering Systems/
Work Area	Functional Analysis and Requirements/ Databases and Data Mining//Infrastructure Development/ Business Processes/ Computer Security/Testing/Several/ Does Not Work
Age	Children under the age of 25 years/ Between 25 and 30 years old / older than 30 years/
Family Group	With/ Without Compromise

The attribute originally had four age ranges (under the age of 25, 25 to 30, 31 to 35, older than 35), regrouped by considering the following ranges (under the age of 25 years, between 25 and 30 years, and older than 30 years) to achieve a more balanced distribution. In the data collection, the Family Group attribute was defined with four values (lives-with-parents, alone, children, and couple). An analysis of this attribute regrouped the data into two values (with compromise, without compromise). This decision is based on the dedication time that the thesis student has for the thesis, the student who lives alone (no children), or with parents, have more time compared to those who are married or living with a partner and/or with children [9] [10].

3. Results

This study describes a preliminary model of descriptive nature in which the aim is not to predict new data, but to describe the existing ones [11]. This model allows to identify the research areas of the thesis selected by the thesis students and its relationship with other attributes that define it. For the construction of the model, a descriptive task was performed [12] for discovering the association rules among the attributes (career, age, family group, working area) and the target attribute (thesis area). The decision tree algorithms were used for experimentation: J4.8 of WEKA [13], ID3 in RapidMiner [14] and CART IN Knime [15].

3.1 Construction of the preliminary model

The decision tree algorithms used for the construction of the model were J4.8 (WEKA), ID3 (in RapidMiner), and CART (in Knime). The attributes are described in Table 1. The thesis area was selected as the meta attribute or class for working without pruning. From the construction of the first model in the three tools, the following sizes of trees were observed: CART (59), (45) J4.8, ID3 (54) [16], in addition to the complex visualization and interpretation for the solution of the problem. The Work Area attribute is used as root node by the three algorithms. The attribute Career in J48 and ID3 is the second node chosen, while in CART it is used in the third or fourth division lines. The data set was divided into two groups after the checks and according to the decision of investigating the problem through classification methods, and with the interest of determining if there are differences between both careers in the elections of the thesis students. The first group for Systems, and the second one for Informatics Engineering.

```
Área = Anal-func-Requ: Ing-Soft {Ing-Soft=6, Ag-Sis-Int=0, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-Aeducacion=0, Arq-
RedesySO=0, Proc-Señales-STR=0}
Área = Seg-Inf: Arq-RedesySO {Ing-Soft=0, Ag-Sis-Int=0, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-Aeducacion=0, Arq-
RedesySO=2, Proc-Señales-STR=0}
Área = desarrollo
| edad = 25-30: Ag-Sis-Int {Ing-Soft=0, Ag-Sis-Int=2, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-Aeducacion=0, Arq-
RedesySO=0, Proc-Señales-STR=0}
| edad = mayor30
| | Grupo-familiar = con-compromiso: Ag-Sis-Int {Ing-Soft=0, Ag-Sis-Int=2, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-
Aeducacion=0, Arq-RedesySO=0, Proc-Señales-STR=0}
| | Grupo-familiar = sin-compromiso: Proc-Señales-STR {Ing-Soft=0, Ag-Sis-Int=0, In-Sis-Soft=0, Seg-Inf=0, Tec- Inf-
Aeducacion=0, Arq-RedesySO=0, Proc-Señales-STR=2}
| edad = menor-25: Ag-Sis-Int {Ing-Soft=0, Ag-Sis-Int=2, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-Aeducacion=0, Arq-
RedesySO=0, Proc-Señales-STR=0}
Área = infraestructura: Seg-Inf {Ing-Soft=0, Ag-Sis-Int=0, In-Sis-Soft=0, Seg-Inf=2, Tec-Inf-Aeducacion=0, Arq-
RedesySO=0, Proc-Señales-STR=0}
Área = no
| edad = 25-30: Ag-Sis-Int {Ing-Soft=0, Ag-Sis-Int=2, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-Aeducacion=0, Arq-
RedesySO=0, Proc-Señales-STR=0}
| edad = menor-25: Tec-Inf-Aeducacion {Ing-Soft=0, Ag-Sis-Int=0, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-Aeducacion=2,
Arq-RedesySO=0, Proc-Señales-STR=0}
Área = testing: Ing-Soft {Ing-Soft=2, Ag-Sis-Int=0, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-Aeducacion=0, Arq-RedesySO=0,
Proc-Señales-STR=0}
Área = varios
| edad = 25-30
| | Grupo-familiar = con-compromiso: Proc-Señales-STR {Ing-Soft=0, Ag-Sis-Int=2, In-Sis-Soft=0, Seg-Inf=0, Tec- Inf-
Aeducacion=0, Arq-RedesySO=0, Proc-Señales-STR=2}
| edad = mayor30
| | Grupo-familiar = con-compromiso: In-Sis-Soft {Ing-Soft=2, Ag-Sis-Int=2, In-Sis-Soft=4, Seg-Inf=2, Tec-Inf-
Aeducacion=0, Arq-RedesySO=0, Proc-Señales-STR=0}
| | Grupo-familiar = sin-compromiso: In-Sis-Soft {Ing-Soft=0, Ag-Sis-Int=2, In-Sis-Soft=2, Seg-Inf=0, Tec-Inf-
Aeducacion=0, Arq-RedesySO=0, Proc-Señales-STR=0}
| edad = menor-25
| | Grupo-familiar = sin-compromiso: Ing-Soft {Ing-Soft=2, Ag-Sis-Int=2, In-Sis-Soft=0, Seg-Inf=0, Tec-Inf-
Aeducacion=0, Arq-RedesySO=0, Proc-Señales-STR=0}
```

Figure 1. A model for Systems using the ID3 algorithm using RapidMiner [17].

Although the J4.8 algorithm in the WEKA tool is clearer to interpret, it does not correctly classify all instances. For example: Area=Several, age=25- 30, Proc-s-Str (4.0/2.0) classifies 4 instances of which 2 are incorrect.

In CART, the rule is more complex as it performs binary divisions and, in the case of attributes with multiple values, it is difficult to read. However, between the ages 25-30 in those working in several areas, it only classifies the instances of Agents and Intelligent Systems, leaving unclassified instances of Signal Processing. It is concluded that this algorithm is not recommended for attributes with multiple values [18].

3.2 Evaluation Phase and Interpretation

Table 2 presents a comparative analysis of the percentage that represents the work areas with greater incidence in the selection of the thesis area.

This table displays the most selected thesis area Agents and Intelligent Systems for both careers. The thesis students were working on several areas at the beginning of the thesis. The obtained decision trees confirmed that Intelligent Systems is the most selected area with a non-significant percentage distribution for both careers.

Table 2. Comparison of the incidence of the work area attribute in the selection of the thesis area (both careers).

	Total, students	**Engineering**	**Bachelor's degree**
Innovation in Software Systems	14%	17%	16%
Technology and Education	15,3%	22%	
Agents and Intelligent Systems	42.51%	47%	32%
Software Engineering	11%		24%
Total of 3 areas	84%	81.3%	75%
Work Area Development	18.54%	11%	12%
Several	49%	44.2%	47%
Total of 2 work area	65.52%	58%	62%

The general trends found in evaluating the decision trees are the following:

-About Systems, the obtained decision trees show that people who worked in the classification Several are not interested in the thesis area: Agents and Intelligent Systems. Majority tendency among all thesis students. On the contrary, it is selected by the thesis students belonging to the area of Development. The thesis students who worked as Functional Analysts, tend to select the Software Engineering area. All the thesis students who select Innovation in Software Systems worked in Several.

-About Informatics Engineering, the Agents and Intelligent Systems thesis area is selected by both students working in Development and the ones who selected Several. The thesis students working in Development are also interested for Innovation in Software Systems. Those who selected Technology and Education, worked in the following areas: Functional Analyst, Infrastructure, and Security in Computer Science.

The resulting models for each of the careers (Systems and Engineering in Computer Science) formalize the reality perceived by teachers of the thesis chairs. However, in order to cover a wider area of reality related to selecting the research area, it is important to consider to widen the quantity of attributes that describe the thesis students [19].

4. Conclusion

A methodological process suggested in KDD was applied to solve the problems found with respect to the data and the construction of the models, which allowed to check the interactive nature of the process. The thesis students' profiles were described for providing useful information in relation to the incidence of the work area at the beginning of the thesis and the research area selected for the thesis. The identified future studies are: (a) the refinement of models achieved by adding new attributes for subsequent analysis on the incidence in the solution of the problem; (b) obtaining new models by using DM techniques to allow the determination of the variables that affect the selection of the thesis area.

References

[1] Ebrahimzadeh, I., Shahraki, A., Shahnaz, A. y Myandoab, A. (2016) Progressing urban development and life quality simultaneously. City, Culture and Society 7, (3), 186-193. 9.

[2] C. Vásquez, M. Torres-Samuel, A. Viloria, J. Lis-Gutierrez, T. Crissien, N. Valera y D. Cabrera, «Cluster of the Latina American universitiers Top100 according to Webometrics 2017» de Lectur Notes in Computer Sciencie (Including subseries Lectur Notes in Artifitial Intelligent and Lectur Notes in Bioinformatics), 2018.

[3] R. Melero y F. Abad, «Revistas Open Access: Características, modelos económicos y tendencias,» Lámpsakos, pp. 12-23, 2001.

[4] M. Pinto, J. C. J. Alonso, V. Fernández, C. García, J. Garía, C. Gómez, F. Zazo y A.-V. Doucet, «Análisis cualitativo de la visibilidad de la investigación en las Universidaes españolas a través de su página Web,» Rev. Esp. Doc., pp. 345-370, 2004.

[5] M. Torres-Samuel, Vásquez, C, A. Viloria, J. Lis-Gutierrez, T. Crissien y N. Valera, «Web visibility profiles of Top100 Latin American Universities, » de Lectur Notes in Computer Sciencie (Including subseries Lectur Notes in Artificial Intelligent and Lectur Notes of Bioinformatics, 2018.

[6] ARWU, «Ranking de Shanghái,» [En línea]. Available: http://www.shangairanking.com/ARWU2018.html. [Último acceso: 01 12 2018].

[7] QS, «QS World University Ranking» [En línea]. Available: https://www.topuniversities.com/university-rankings/latin-american-university-rankings/2019. [Último acceso: 01 12 2018].

[8] Scimagooir, «Scimago,» Scimago Lab, [En línea]. Available: https://www.scimagoir.com/methodology.php. [Último acceso: 01 06 2019].

[9] M. Torres-Samuel, C. Vásquez, A. Viloria, L. Hernández-Fernandez y R. Portillo-Medina, «Analysis of patterns in the university Word Rankings Webometrics, Shangai, QS and SIR-Scimago: case Latin American» de Lectur Notes in Computer Science (Including subseries Lecture Notes in Artificial Intelligent and Lecture Notes in Bioinformatics, 2018..

[10] Torres-Samuel, M., Vásquez, C., Viloria, A., Lis-Gutiérrez, J.P., Borrero, T.C., Varela, N.: Web Visibility Profiles of Top100 Latin American Universities. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, Springer, Cham, vol 10943, 1-12 (2018).

[11] Viloria, A., Lis-Gutiérrez, J.P., Gaitán-Angulo, M., Godoy, A.R.M., Moreno, G.C., Kamatkar, S.J. : Methodology for the Design of a Student Pattern Recognition Tool to Facilitate the Teaching – Learning Process Through Knowledge Data Discovery (Big Data). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data.

[12] C. Vásquez, M. Torres-Samuel, A. Viloria, J. Lis-Gutierrez, T. Crissien, N. Valera, D. Cabrera y M. Gaitán-Angulo, «Efficiency analysis of the visibility of Latin American universities and theri impact on the Ranking Web,» de Lecture Notes in Computer Science (Including subseries Lectur Notes in Artificial Intelligent and Lecture Notes in Bioinformatics), 2018

[13] Sekmen, F., Kurkcu, M.: An Early Warning System for Turkey: The Forecasting of Economic Crisis by Using the Artificial Neural Networks. Asian Economic and Financial Review 4(1), 529-43 (2014).

[14] A Lee, P Taylor, J Kalpathy-Cramer, A Tufail Machine learning has arrived!. Ophthalmology, 124 (2017), pp. 1726-1728

[15] Yao L (2006). The present situation and development tendency of higher education quality evaluation in Western Countries. Priv. Educ. Beef. (2006).

[14] Gregorutti B, Michel B, Saint-Pierre P (2015) Grouped variable importance with random forests and application to multiple functional data analysis. Comput Stat Data Anal 90:15–35

[15] Isasi, P., Galván, I.: Redes de Neuronas Artificiales. Un enfoque Práctico. Pearson. ISBN 8420540250 (2004).

[16] Haykin, S.: Neural Networks and Learning Machines. New Jersey, Prentice Hall International (2009).

[17] Zhang, G.P.: Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50 (1), 159-75 (2003).

[18] Kuan, C.M.: Artificial neural networks. In the New Palgrave Dictionary of Economics, ed. S.N. Durlauf and L.E Blume. UK: Palgrave Macmillan (2008).

[19] Pineda Lezama, O., Gómez Dorta, R.: Techniques of multivariate statistical analysis: An application for the Honduran banking sector. Innovate: Journal of Science and Technology, 5 (2), 61-75 (2017).