redin Revista Facultad de Ingeniería, Universidad de Antioquia, No.87, pp. 70-76, 2018

Conceptual clustering: a new approach to student modeling in Intelligent Tutoring Systems

Agrupamiento conceptual: un nuevo enfoque para el modelado del estudiante en Sistemas Tutoriales Inteligentes

Yunia Reyes-González*, Natalia Martínez-Sánchez[1], Adolfo Díaz-Sardiñas[1], Marisol de la Caridad Patterson-Peña[1]

[1]Departamento de Programación, Universidad de las Ciencias Informáticas (UCI). Carretera de San Antonio de los Baños, Km. 2 ½, Comunidad Torrens, Municipio Boyeros. C. P. 19370. La Habana, Cuba.

ARTICLE INFO:
Received April 04, 2017
Accepted April 18, 2018

KEYWORDS:
Student modeling, intelligent tutoring systems, logical combinatorial pattern recognition, artificial intelligence

Modelado del estudiante, sistemas tutoriales inteligentes, reconocimiento lógico combinatorio de patrones, inteligencia artificial

ABSTRACT: Student modeling is a central problem in Intelligent Tutoring Systems design and development. In this way, the characteristic that distinguishes this type of system is the ability to determine as accurately and quickly as possible the student's cognitive and affective-motivational state in order to personalize the educational process. Therefore, the fundamental problem is to select data structure to represent all relative information to student and to choose the procedure to make the diagnosis. This paper describes a model for knowledge engineering inherent to all intelligent tutoring system, using the LC-Conceptual clustering algorithm, from logical combinatorial pattern recognition. This algorithm builds the objects clusters based on their similarity, using a grouping criterion, and it also builds the property (or concept) that meets each group of objects.

RESUMEN: El modelado del estudiante es un problema central en el diseño y desarrollo de los Sistemas Tutoriales Inteligentes. En este sentido, la característica que distingue este tipo de sistema es la capacidad de determinar con la mayor precisión y rapidez posible cuál es el estado cognitivo y afectivo-motivacional del estudiante para personalizar el proceso de enseñanza-aprendizaje. Por lo tanto, el problema fundamental está en seleccionar la estructura de datos para representar toda la información relativa al estudiante y elegir el procedimiento para realizar el diagnóstico. En este trabajo se describe un modelo para realizar la ingeniería del conocimiento inherente en todo Sistema Tutorial Inteligente utilizando el algoritmo de agrupamiento LC-conceptual del reconocimiento lógico combinatorio de patrones el cual además de construir los agrupamientos de objetos, basándose en la semejanza entre los mismos y utilizando un criterio de agrupamiento, construye la propiedad (el concepto) que cumple cada agrupamiento de objetos.

1. Introduction

Intelligent Tutoring Systems (ITS) are programs to manage knowledge about a certain area or subject with the purpose of transmitting this knowledge to the students through an individualized interactive process. ITS also try to simulate the way a tutor guides the students through the teaching-learning process. "Intelligent" refers to the system ability on what to teach, when to teach and how to teach, simulating the role of a real teacher. To achieve this goal, ITS should find relevant information about the students learning process and apply the best instructional way, according to their individual needs [1].

ITS [2] are divided into three basic modules. The Student Module stores the students' characteristics according to their knowledge level. This module determines what the students know and from this point, it can be inferred what to teach and how to teach it. This information is represented in the Domain Module and the Training Module.

The knowledge-based systems [3] are valid artificial intelligence techniques to build ITS, due to their related approaches. Those systems use specific domain knowledge. The solution obtained is similar to the one given by a skilled person in the problem domain. ITS use stored information about the student characteristics, to adapt the didactic material to the subject that is taught.

* Corresponding author: Yunia Reyes González
E-mail: yrglez@uci.cu

ISSN 0120-6230
e-ISSN 2422-2844

DOI: 10.17533/udea.redin.n87a09

The current trend is aimed at using artificial intelligence techniques in the implementation of ITS in some of their modules. ITS use artificial intelligence techniques in the Domain Model, such as: frames and Bayesian networks. The Educational module implementation uses plans, production rules, neural networks, semantic networks and Bayesian networks. In other developments, the approach is to develop the student model using techniques such as: Bayesian networks; Bayesian Markov Chain Clustering [4], ontologies; fuzzy logic; machine learning [5]; and case-based reasoning [6].

The logical combinatorial pattern recognition [7] can be applied to solve problems in most knowledge areas such as: character recognition, medical diagnosis, remote sensing of the earth, human face identification and fingerprints, forecast breaks in machinery and equipment, signal analysis and biomedical imaging, automatic inspection, blood count, archeology, mineral deposit forecasting, analysis of seismic activity and document classification. Early applications are founded on [3–8] and specifically conceptual grouping on [9].

In particular, it is clear that ITS development requires not only application domain knowledge, but also programming and artificial intelligence skills, which are difficult for one person to have them all together. According to this, the development of such systems is possible only with a multidisciplinary approach.

This research begins with the first ideas published in [9], and it is based on the facilities that conceptual clustering algorithms offer to modeling student implementations. This allows structuring the universe of knowledge and finding out the meaning of each structure qualitatively. This conceptual clustering provides the concept which is implicit around the properties met by the contained objects in every structure, in such a way that it reflects qualitatively what the specialist handles within the area and context of the problem to be solved.

2. Intelligent Tutoring System Modules

The architecture presented in [10] brings together the elements most commonly used, and it is summarized in the statement that ITS are composed by a domain module, a student module and an educational module. They are communicated interactively through a central module that is often called environment module.

In fact, the student model is a research problem that should be approached from all edges in order to obtain a complete and accurate representation of the students' knowledge. Some authors take into consideration features such as learning style, knowledge level, personal information or their combination.

The domain module, also called expert module by many authors, provides the knowledge domain and fulfills two different purposes:

- First, to present the subject in an appropriate way for students to acquire the skills and concepts, including the ability to generate questions, explanations, answers and tasks.

- Second, the domain module should be able to address problems, to correct solutions and to accept any valid solutions which have been obtained by different means. In this module, the knowledge to be taught should be didactically organized to facilitate the teaching-learning process.

Additionally, the pedagogical module decides what, how and when to teach the tutor contents, adapting their teaching decisions to students' needs. Some authors refer to such module as a tutor module, since it is responsible for comparing the students' characteristics with teaching content and choosing the best way to make appropriate instructional decisions.

3. Students modeling using conceptual clustering

Student modeling is the implicit inference method in all ITS to diagnose cognitive-affective student state. As discussed in [8], ITS are knowledge-based systems that decide what to teach and how to develop the teaching-learning process of a new student through accumulated experiences in the student module. Therefore, in the knowledge-engineering process, it is necessary to determine the behavior of stored models in the student module, so that this analysis helps to elaborate the teaching materials adapted to their distinctive characteristics.

All this consideration implies the treatment of an unsupervised problem, where it is essential to group the students' models according to their degree of similarity and to determine what features distinguish them.

Almost all algorithm models for unsupervised problems provide structuring of the spaces, on which they are applied extensionally, i.e., they determine which objects are in a certain grouping, or what degree belongs to a grouping. In other words, they are given an extensional structure of space.

However, Michalski's conceptual model was an exception.

His algorithm model determines not only what the integration of their groupings is, but also what properties meet the objects belonging to the same grouping. It can be stated that Michalski's algorithms were the first to provide a conceptual structure of space [7].

Conceptual clustering algorithms can be divided into two groups, incremental and non-incremental algorithms. Incremental algorithms base their operation on the adaptation of the groupings (or concepts) with new objects that are being presented; that is, whenever a new object is given using a certain strategy, it is classified into existing clusters or new groupings are created. On the other hand, non-incremental algorithms structure a sample of objects without assuming that they are given one by one.

The model described in this paper is based on the essential idea of the LC-Conceptual algorithm [11]. This algorithm includes two phases: the first one is the extensional structure, where groups of objects are built based on their similarity and using a grouping criterion. In the intentional structure phase, the property (or concept) that each group of objects meets is constructed.

As a result of the extensional stage, student model groups are obtained. On the intentional stage, concepts associated with each group are generated; these concepts are characterized by having no objects with similar properties in other groups.

An advantage of this stage is that concepts that do not describe group objects are obtained (unobservable objects, see [12]) although these objects are similar to the ones within the same group. This is very useful during the knowledge engineering phase, as the concepts can be taken into consideration for the development of teaching materials, after a preliminary analysis with the experts where ITS are being developed.

Following, there is a description in pseudo-code of the algorithms included in the proposed model for students modeling using the basic ideas of LC-Conceptual algorithm [11].

The model consists of two phases: the first one for knowledge engineering and the second phase for student modeling using concepts. In the first stage, the ITS are running, and given a new student model they determine what to teach and how to develop the teaching-learning process.

3.1 Phase 1: Students models characterization

The initial matrix (MI) is taken as input, Equation. (1), where n is the number of features (characteristics that describe student model) and m is the number of objects (model students) which is included in the matrix (Student Module). Algorithm 1, for organizing students in similar groups, is shown in Algorithm 1.

$$\begin{array}{c|ccc} & X_1 \ldots & & X_n \\ O_1 & X_1(O_1) & \cdots & X_n(O_1) \\ \vdots & \vdots & \vdots & \vdots \\ O_m & X_1(O_m) & \cdots & X_n(O_m) \end{array} \quad (1)$$

```
input  : MI                     // Student module
output:
Training Matrix (TM)      // groups are formed
according to similarity among the
student's models.
CA     // set of concepts associated with
each group.
Step 1 Extensional stage of LC-Conceptual    // TA
 is obtained
Step 2 Calculate typical testors using FastBR
 algorithm [13]  // They are calculated from
 the TM and a set of typical testors
 (TT) is obtained.
```

Algorithm 1: Organization of students in similar groups

As the number of typical testors can be large, ordering is required by utility level. Several concepts can be generated out of a typical testor; they may be sufficient to characterize the students' models that comprise them, or it may be necessary to calculate the concepts from several typical testors. Algorithm 2 illustrates algorithm for calculating the testors usefulness.

3.2 Phase 2: Student Modeling

This phase implements the student modeling. Given a new student model, it is situated in the most similar group according to its characteristics. This allows adapting students features to didactic materials associated with the group. Algorithm 3, shows how to perform this process.

```
input  : TT
output: TT'                                            // Set of more useful testors
Step 1 Calculate the weight of $\epsilon_i$ of features $x_i$ appearing in the testors' family as Equation. (2):

    $\varepsilon_i(x_i) = \alpha P_i(x_i) + \beta L_i(x_i)$                                   (2)

  where: $\alpha > 0$, $\beta > 0$ and $\alpha + \beta = 1$. $\alpha$ and $\beta$ are two parameters that weigh the influence of $P_i$ and $L_i$
  (frequency of occurrence as Equation. (3) and length of testors as Equation. (4)).

    $P(x_i) = \frac{|T_i|}{|\mathrm{T}|}$                                                    (3)

    $L(x_i) = \frac{\sum_{t \in T_i} \frac{1}{|t|}}{|T_i|}$                                   (4)

  where: $|T_i|$ number of testors where feature $i$ appear $|\mathrm{T}|$ number of testors $|t|$ number of features that form the
  $T_i$ Testor
Step 2 Selecting the set of relevant features: For each testor ti calculate Equation. (5):

    $\psi_i(t_i) = \sum_{i=1}^{|t_i|} \varepsilon(x_i)$                                       (5)

   // the calculated amount is the average of the importance of features that make up the
  testor.
Step 3 Select the p testors higher $\psi_i(t_i)$ where $p$ is a parameter related to the problem to be solved and $TT'$ is
  the set of $p$ testors.
Step 4 Intentional stage of LC-Conceptual algorithm. // The concepts associated with each group are
  calculated from TM and TT'.
```

Algorithm 2: Algorithm for calculating the testors usefulness[9]

```
input  :
TM                 // groups formed according to similarity between the students' models
CA                                        // set of concepts associated with each class
$O_t$                                                          // new student model
B                                                            // similarity function
output:
$G_i$                                     // Group of most similar students according to $O_t$
Step 1 For each $G_i$ of TA calculate:                                       // $G_i$: group $i$
a. $\{B_i(O_t, O_j)\}$    // where $O_j$ students are models that correspond to the CA'. It is valid
  to clarify that only the features that make the CA' are compared.
b. Calculate Equation. (6):
    $\lambda_i(O_t) = \sum \beta(O_t, O_l) \, / \, |l|$                                     (6)
    // $O_t$ is the new student model, $O_l$ concepts and $|l|$ the number of concepts associated
    with the group $i$.
Step 2 a. Select as Equation. (7), the max
    $(\lambda_i(O_t))$                                                                    (7)
  b. Select the group $G_i$  // the objects of group $i$ correspond to student's models who make up
  the group $i$, more similar to the new student model.
```

Algorithm 3: Adapting the didactic material according to students' characteristics

Throughout Example 1 the two phases of the proposed model are illustrated.

Example 1. Given the array MI (student module) where O_i represent eight students' models and r_i the five features that describe them. The features r_1, r_2, r_3 can take as values: {A, B, C, D, E}, and r_4, r_5 can take as values {1, 0}. When Phase 1 is completed, the TA matrix composed by two groups {I, II} is obtained. Table 1 shows the initial matrix, and Tables 2 and 3 show two groups making up the training matrix. The concepts covering the observable and unobservable objects are shown, too. For each group unobservable concepts are underlined.

Table 1 Initial Matrix

Objects/Features	r_1	r_2	r_3	r_4	r_5	Group
O_1	A	C	B	0	1	¿?
O_2	B	A	A	1	1	¿?
O_3	A	B	B	0	1	¿?
O_4	A	A	A	0	1	¿?
O_5	C	E	D	1	0	¿?
O_6	D	D	D	1	1	¿?
O_7	C	E	D	0	0	¿?
O_8	A	D	A	1	1	¿?

Suppose Phase 1 is completed, two groups I and II are formed and the most useful typical testor is made up by the features: {r_1, r_4,}.

Table 2 Objects of Group I

Objects/Features	r_1	r_2	r_3	r_4	r_5
O_1	A	C	B	0	1
O_2	B	A	A	1	1
O_3	A	B	B	0	1
O_4	A	A	A	0	1

From Table 2, corresponding to objects of Group I, the possible concepts are: {A,0}; {B,1};{B,0};{A,1}.

Table 3 Objects of Group II

Objects/Features	r_1	r_2	r_3	r_4	r_5
O_5	C	E	D	1	0
O_6	D	D	D	1	1
O_7	C	E	D	0	0
O_8	A	D	A	1	1

According to Table 3, for objects of Group II, the possible concepts are: {C,1}; {D,1}; {C,0}; {A,1}; {D, 0}; {A,0}.

It is important to note that the possible concept ~~{A,1}~~ of Group I is eliminated because it appears in observable concepts of Group II. Furthermore, {D, 0} is an unobservable concept of Group II because its characteristics are typical of this group, but no object in this group fulfills them. The latter is the case in which the analysis with experts in the knowledge area is recommended, as this concept may not necessarily occur.

Phase 2 is described from Example 2.

Example 2: Given a new student model: $O_{new} = \{B, E, A, 1, 1\}$ the purpose is to determine the educational material needed to develop a customized educational process.

Example 2.1 The new student model t: $O_{new} = \{B, E, A, 1, 1\}$ is compared with the concepts of Group I, and as consistent with the concept {B,-,-,1,-} the student develops the teaching learning process with the teaching material associated to group I.

Example 2.2 $O_{new} = \{D, B, D, 0, 0\}$, is compared with Group I concepts, but there are mismatches. Then, it is compared with Group II concepts and this case matches with concept {D,-, -,0,-}. It is an unobservable concept because it does not represent any object in the group. The procedure continues as explained before.

4. Students modeling validation using the LC-Conceptual algorithm

For the validation of the proposed model, the ITS for studying the basic theory of Logical Combinatorial Pattern Recognition, was used [8]. This system uses the CLASS algorithm [14] from logical combinatorial pattern recognition for clustering (Phase 1 in the proposed model). The holotype of each group is calculated for students modeling only in the initialization stage. The holotype of a group [14] is the object that most closely resembles the other objects within the group. That is, in phase 2 the new model is compared to the holotypes of each group, and the new students can study with the educational material associated to the group whose resemblance to the holotype turned out to be higher.

Table 4 shows the comparison of ITS with three different models for studying the basic theory of Logical Combinatorial Pattern Recognition using different methods of knowledge engineering and student modeling.

The expert method described in [15] was used in order to evaluate the model feasibility. Variables were weighed considering the satisfaction frequency and data descriptive analysis was performed using the SPSS statistical package. Firstly, the number of experts was defined

Table 4 Comparing ITS with two different models

Comparative aspects	ITS using method [8]	ITS using the method [9]	ITS using the model proposed in this research
Number of groups according to the similarity among the students' models stored in students module	EQUAL		
Feasibility models	4.36	4.44	4.65
Students modelling efficiency	96.2 %	100 %	

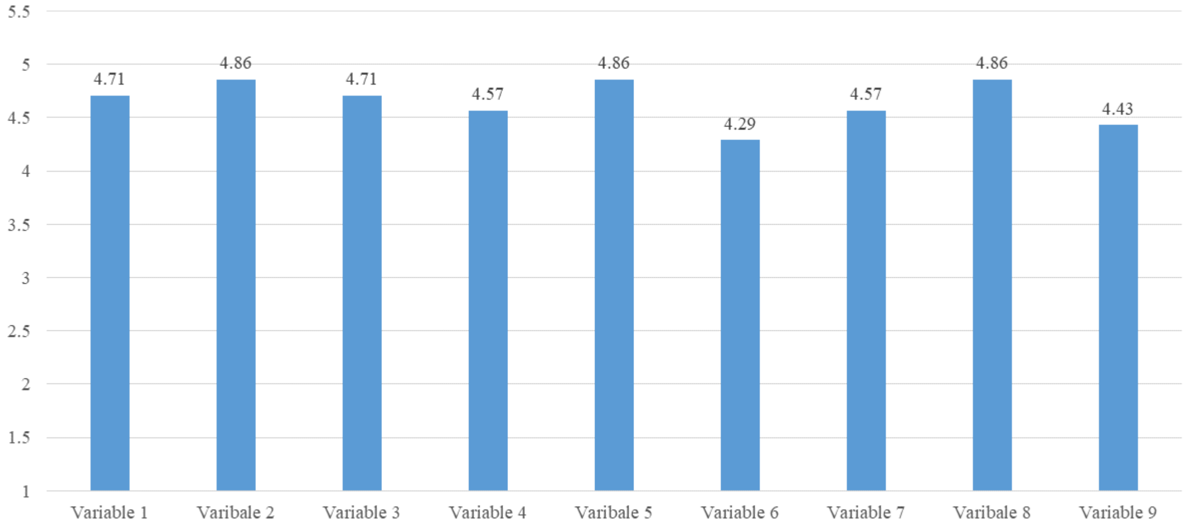

Figure 1 Results in experts' evaluation

through a probabilistic method and assuming a binomial probability law, with 10% of precision, 1% of errors estimated proportion (average) and a confidence level of 99%. Seven (7) experts were chosen for the study.

The experts were selected according to their knowledge and professional experience on the topic of study. Next, the variables listed below were defined to evaluate the model feasibility:

Variable 1 The model allows forming groups of students with similar characteristics.

Variable 2 The model enables the identification of the most important characteristics.

Variable 3 The concepts of each group constitute distinctive student's characteristics of that group in respect to the rest of the groups.

Variable 4 The model allows personalizing the assigned didactic material process to the students since it adapts to their characteristics.

Variable 5 The model allows classifying effectively new students in previously formed classes, taking into account the concepts of each class.

Variable 6 The model guarantees an adequate representation of the information related to the student cognitive state.

Variable 7 The use of conceptual algorithms provides added value to the knowledge engineering stage in designing ITS.

Variable 8 The model is relevant and applicable in the current educational context.

Variable 9 The model is useful for the topic selections that make it possible to describe the students cognitive state.

The experts evaluated each variable on a Likert scale [16], where 5 represents the expert's complete agreement and 1 the means the expert totally disagrees with the evaluated variable. All the average values resulting from the expert's evaluations exceed the value of 4; therefore, the application of the model is feasible, as shown in Figure 1.

Finally, the expert variation coefficient of each variable was calculated, and this was lower than 0.20, subsequently there was agreement among the experts.

To validate the student modeling efficiency, k-fold Cross Validation method for k = 10 was carried out. Given a new student model, it was correctly classified and thus a didactic material was assigned in order to customize this process.

From a research conducted in [17], classic inference measures from rough sets theory are used to compare the model proposed in [9] and the model proposed in this research. This shows that typical testors selection and the inclusion of unobservable objects concepts ensures full coverage of the objects (students models) from the universe (Student Module). This is shown in results obtained by calculating the measures as Equation. (8) and Equation. (9):

$$\alpha\left(x\right)=\frac{\left|R_{8}^{'(x)}\right|}{\left|R'^{*}\left(x\right)\right|}=1 \tag{8}$$

$$\gamma\left(x\right)=\frac{\left|R_{*}^{'(x)}\right|}{\left|x\right|}=1 \tag{9}$$

5. Conclusions

As a result of this research, a model is obtained, which can be considered in the ITS development. It offers a new vision for the students modeling using the basic ideas of conceptual clustering algorithms from the logical combinatorial pattern recognition.

The proposed model provides a feasible and effective method for the knowledge engineering stage as shown in the validation performed. Students' Models are grouped in classes (or clusters) according to their degree of similarity and the distinctive features (concepts) that characterize the clusters are determined. The concepts meet the property to ensure there are no student models in other groups with the same characteristics, and they cover the entire universe of objects. According to the authors of this research, this is the most valuable feature of the proposed model.

Having these features allow the development of didactic materials that make up the ITS with a high level of customization. In addition, the validation performed shows effectiveness in student modeling to correctly classify 100% of the selected students' models.

The proposed algorithms are computationally expensive (because of the algorithms for the calculation of the typical testors and the operator to calculate the concepts, both of exponential cost as referred to in the literature revised [7]), but they are applied only during the knowledge engineering stage as part of the whole work implied in ITS development such as: diagnostic context, implicit methodological work, topic selection, structuring and definition of the pursued objectives. These particularities are among others the basis for the students' characterization and the appropriate instructional materials for each student model.

References

[1] D. A. Ovalle and J. A. Jiménez, "Entorno Integrado de Enseñanza / Aprendizaje basado en Sistemas Tutoriales Inteligentes Ambientes Colaborativos," *Sistemas, Cibernética e Informática*, vol. 1, no. 1, pp. 23–27, 2004.

[2] N. Martínez, M. M. García, and Z. Z. García, "Modelo para diseñar sistemas de enseñanza-aprendizaje inteligentes utilizando el razonamiento basado en casos," *Revista Avances en Sistemas e Informática*, vol. 6, no. 3, pp. 67–78, Dec. 2009.

[3] T. J. M. Bench, *Knowledge Representation: An Approach to Artificial Intelligence*, 1st ed. San Diego,USA: Academic Press, 1990.

[4] C. Li and J. Yoo, "Modeling Student Online Learning Using Clustering," in *44th annual Southeast regional conference*, Melbourne, Florida, 2006, pp. 186–191.

[5] K. Chrysafiadi and M. Virvou, "Student modeling approaches: A literature review for the last decade," *Expert Syst. Appl.*, vol. 40, no. 11, pp. 4715–4729, Sep. 2013.

[6] D. Medina, N. Martínez, Z. Z. García, M. Chávez, and M. M. García, "Putting Artificial Intelligence Techniques into a Concept Map to Build Educational Tools," in *Nature Inspired Problem-Solving Methods in Knowledge Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 617–627.

[7] J. Shulcloper, "Reconocimiento lógico combinatorio de patrones: teoría y aplicaciones," M.S. thesis, Universidad Central de Las Villas, Santa Clara, Cuba, 2009.

[8] N. Martinez, M. M. Garcia, and J. E. Hurtado, "Model for designing Intelligent Tutorials Systems using Conceptual Maps and knowledge-based Systems," *IEEE Latin America Transactions*, vol. 10, no. 6, pp. 2301–2308, Dec. 2012.

[9] Y. Reyes and N. Martínez, "La toma de decisiones en los Sistemas Tutoriales Inteligentes utilizando el agrupamiento conceptual," *Rev. Cuba. Cienc. Informáticas*, vol. 8, pp. 104–116., Dec. 2014.

[10] D. Ovalle, "Análisis funcional de la estrategia de aprendizaje individualizado adaptativo," Proy. Investig., Universidad Nacional de Colombia, Medellín, 2007.

[11] J. F. Martínez, "Herramientas para la Estructuración Conceptual de Espacios," M.S. thesis, CIC, IPN, México, 2000.

[12] R. S. Michalski, "Conceptual Clustering: A Theoretical Foundation and a Method for Partitioning Data into Conjunctive Concepts," in *Textes des exposes du Seminaire organise par l'Institute de Recherche d'Informatique et d'Automatique (IRIA)*, París, France, 1979, pp. 254–294.

[13] A. Rodriguez and G. Sánchez, "An Algorithm for Computing Typical Testors Based on Elimination of Gaps and Reduction of columns," *Aerosp. and Electron. Syst.*, vol. 27, no. 8, p. 18, Dec. 2013.

[14] J. Martínez and A. Guzmán, "The logical combinatorial approach to pattern recognition, an overview through selected works," *Pattern Recognit.*, vol. 34, no. 4, pp. 741–751, Apr. 2001.

[15] J. Plasencia, F. Marrero, M. Nicado, and Y. Aguilera, "Procedimiento para la priorización de Factores Críticos de Éxito," *DYNA*, vol. 84, no. 202, pp. 26–34, Jul. 2017.

[16] R. Likert, "A technique for the measurement of attitudes," *Arch. Psychol.*, vol. 22, no. 140, pp. 5–55.

[17] Y. Reyes, N. Martínez, and M. M. García, "El agrupamiento conceptual en el contexto de la teoría de los conjuntos Aproximados," *DYNA New Technol.*, vol. 2, no. 1, p. 12, Jan. 2015.