JOURNAL OF INFORMATION SYSTEMS ENGINEERING & MANAGEMENT

Journal of Information Systems Engineering
& Management, 2018, 3(1), 06
ISSN: 2468-4376

lectito

Identifying User Profiles from Statistical Grouping Methods

Francisco Kelsen de Oliveira [1, 2*], Max Brandão de Oliveira [3], Alex Sandro Gomes [2],
Leandro Marques Queiros [2]

[1] *Instituto Federal de Educação, Ciência e Tecnologia do Sertão Pernambucano (IF Sertão-PE), Pernambuco*, BRAZIL
[2] *Universidade Federal do Pernambuco (UFPE), Pernambuco*, BRAZIL
[3] *Universidade Federal do Piauí (UFPI), Piauí*, BRAZIL

*__Corresponding Author:__ francisco.oliveira@ifsertao-pe.edu.br

Citation: Oliveira, F. K., Oliveira, M. B., Gomes, A. S. and Queiros, L. M. (2018). Identifying User Profiles from Statistical Grouping Methods. *Journal of Information Systems Engineering & Management*, 3(1), 06. https://doi.org/10.20897/jisem.201806

Published: February 10, 2018

ABSTRACT

This research aimed to group users into subgroups according to their levels of knowledge about technology. Statistical hierarchical and non-hierarchical clustering methods were studied, compared and used in the creations of the subgroups from the similarities of the skill levels with these users' technology. The research sample consisted of teachers who answered online questionnaires about their skills with the use of software and hardware with educational bias. The statistical methods of grouping were performed and showed the possibilities of groupings of the users. The analyses of these groups allowed to identify the common characteristics among the individuals of each subgroup. Therefore, it was possible to define two subgroups of users, one with skill in technology and another with skill with technology, so that the partial results of the research showed two main algorithms for grouping with 92% similarity in the formation of groups of users with skill with technology and the other with little skill, confirming the accuracy of the techniques of discrimination against individuals.

Keywords: statistical methods, teacher profiles, skills with technology, user grouping, similarity of profiles

INTRODUCTION

Data collection is an important step in the process of scientific research, which requires selection of the sample of representative form, the selection of appropriate methods, techniques and organization of these data to be analyzed later.

Sampling is one of the first steps of data collection to be set, in order to avoid the bias of the search results and ensure the representation of the populations surveyed, especially, when it is necessary to ensure quotas or percentages by subgroups contained in the sample. The objective of this investigation was to identify statistical methods able to survey individuals group into subgroups with similar features on their skills with technology.

The sample included the participation of teachers who responded to the semi-structured and online survey (Flick, 2013) about the experiences of use of each on aspects and technological tools. The data were organized into tables so that they were run from cluster analysis algorithms (Mingoti, 2005) to assist the definition of groups of users with ability in technology and another without.

In this way, were applied statistical methods to be used for grouping of individuals in the sample, according to the responses submitted to the survey as well as was possible to group individuals into subgroups according to their characteristics.

The results obtained in this research analysis show a likeness of 92% between subgroups formed from the implementation of the algorithms of grouping models with type Ward and K-means. Also that the methods used in this research may cooperate in future investigations that seek to also classify users, define subgroups as features.

THE USERS PROFILE

An important step in the development of any product is to meet the wishes and the needs of the users. This also occurs with the information and communication technology (ICT), so the area of human-computer interaction (HCI) is dedicated to the study and development of more efficient methods and techniques on capturing such data of these users.

The user profile is the individualized description of the characteristics of users (Barbosa and Silva, 2010). Baxter, Courage and Caine (2015) in line with the concept presented and still ensures that the purpose of raising the profile of the user represents really know better for anyone who is developing the product and who will choose to research, validation, satisfaction and other.

Rogers, Sharp and Preece (2013) report that the characteristics of the users should cover the main attributes of the intended user group, highlight the relevant skills and abilities of the user, and even cites some attributes to be considered in the survey of profile: nationality, education, preferences, personal circumstances, physical or mental handicaps and other.

The researches of Courage and Baxter (2005), Hackos and Redish (1998) and Peffer and Renken (2015) describe some types of data for better clarification of the user profile and to be collected for better definition of the domain of the product and the user interface with the technology: demographics, experience in the position he holds, company information, degree of education, experience with computers, experience with specific product or similar tools, available technology, training, attitudes and values, domain knowledge, goals, tasks, severity of errors, motivation to work, languages and jargon.

Highlights the importance of identifying the level of user experience on that if you want to investigate (beginner, expert, casual user or frequent user) as it affects especially the forms of interactions to be designed. This shows the importance of the definition of subgroups of the sample from statistical analysis of the data provided by users in order to truly implement products and their validation mechanisms in accordance with the intended target audience (Rogers et al., 2013).

The researches (Courage and Baxter, 2005) and (Hackos and Redish, 1998) confirm such a point of view by saying that the user profile helps meet to whom the product is being built, as well as collaborate in choosing participants for future activities of analysis and product evaluation.

The data of the users, so they can be collected from interviews and questionnaires. This data will add the values to the groups and tracks which fit together, in order to draw the profiles of users with similar characteristics and set the proportion of users that fit in each profile. It is important to highlight the possibility to prioritize certain features of a user profile, as the product or project in question (Barbosa and Silva, 2010).

In the case of interface design process, users should be identified and characterised from the analysis and modeling of users with the following aspects, according to Oliveira Netto (2010): role or function specific to each user, familiarity with computer, level of knowledge of the field of application, frequency of use of the application and sociocultural context.

Lee (1993) suggests that the analysis of the users can be divided into five steps:

- identification of critical analysis and central factors for implementation;
- Explore other critical factors for implementation;
- Estimate the distribution of users for each factor;
- Identify major groups of users;
- Analyze the collective involvement of the distribution of users.

It is still possible the inclusion of subjective factors on the last item of the classification, it is apparent that the distribution is not a factor. Oliveira Netto (2010) recommends that the questionnaires for analysis of users take into account user's favorite graphical environment (Windows, Linux or MAC OS), frequency of use (occasional, frequent or enumerated amount of times for a period of time) and level of familiarity or expertise in the field of application. In other words, the knowledge of how to perform the same tasks without the aid of the computer (beginner, intermediate or expert).

Thus, this research will prioritize the user experience, while the polling questionnaire also will seek to identify the attitudes of users who can confirm their experiences.

GROUPING METHODS

In this section are presented the main statistical methods, especially, hierarchical types group and non-hierarchical. It is natural to wish to qualify individuals or elements according to a pattern of similarity. The classification is obtained in order to make decisions appropriate for each group in particular, optimizing and directing consistent actions according to the need of the elements that make up each cluster. The assignment of the elements to the groups can be held so subjective, suffering the interference responsible for discriminating against the elements. However, the classification is done impartially, free of human intervention.

A difficulty inherent in the sorting process is associated with the amount of variables under study. The more variables, the more bureaucracy. For both, there are quantitative methods responsible for taking into account the information multivariate sampling unit. To group the elements, analysis decomposes in two instances, the methods of hierarchical and non-hierarchical groups, including details on (Oliveira et al., 2017; RENCHER, 2002).

Hierarchical Methods

Consider *n* elements measured in *p* variables. The formation of groups is associated with responsible for metrics to quantify the similarity or dissimilarity between the observations. The measure most often used in the literature is Euclidean distance, denoted by a quadratic matrix $\boldsymbol{D_n}$ containing 2 to 2 distances between all submissions, number of elements. Each element of $\boldsymbol{D_n}$, d_{ik} is obtained by

$$d_{ik} = \sqrt{(\boldsymbol{x_i} - \boldsymbol{x_k})'(\boldsymbol{x_i} - \boldsymbol{x_k})} = \sqrt{\sum_{j=1}^{p} \left(x_{ij} - x_{kj}\right)^2} \tag{1}$$

where and $\boldsymbol{x_i}$ and $\boldsymbol{x_k}$ are vectors of p dimension regarding elements i and k. The Euclidean distance was employed in the study. This and other measures are presented in (Hardle and Simar, 2007).

The distance is used as a decision criterion in an algorithm whose code is shown in Algorithm 1, which allocates an element to a group to each iteration through a connection method. In this study, the main methods are used: Single, Complete, Average and Ward. The connection methods are discussed in detail in (Johnson and Wichern, 2007).

Algorithm 1: Grouping based on hierarchical method

```
Input: Data $\boldsymbol{A_{nxp}}$ with $n$ measured objects in $p$ variables
Output: Matrix of distances $\boldsymbol{D_{n_xn}}$ among the objects considering all the variables
1.  For $i = 1 \rightarrow n$ then
2.      Method_of_connection ()
3.      Allocate selected object $\boldsymbol{i}$ through the method of connection in a set
4.      Calculate new matrix of distances $\boldsymbol{D_{(n-i)_x(n-i)}}$
5.  End
```

The Single method, also known as nearest-neighbor connection, uses the shortest distance to allocate an element to a group. The Complete, contrary to the Single, groups the neighbors further away and therefore uses the greatest distance between the elements of the group. The Average method is analogous to the previous, except that the distance between the groups is taken as the average between two elements of each group. **Figure 1** illustrates three methods. Finally, in Ward method, the allocation of an element to a group is performed in order to minimize a measure of internal homogeneity, i.e. every step of the method, add objects in order to homogenize the groups. The measure of homogeneity is based on the sum of squares total a analysis of variance (Bieniasz and Majchrzak, 2011).

Non-Hierarchical Methods (K-means)

The method is considered an unsupervised learning algorithm in which the clusters are grouped according to similarity of the elements (Ribas et al., 2012). The method was proposed by Hartigan and Wong (1979) and is still considered one of the most robust clustering algorithms, due to sensitivity to the presence of outliers. For Everitt and Hothorn (2011), the sensitivity is from matrix of distances used in getting the centroides. The number of groups to be composed is indicated by the letter k. In this study, k refers to the ability of the user to operate the software.

The algorithm performed by k-means is to position the clusters in the same space using the Euclidean distance with measurement of similarity. The position of the groups is obtained through the centroid, defined as the sum of each dimension of the space divided by the number of cluster elements. To add an element to the cluster, select

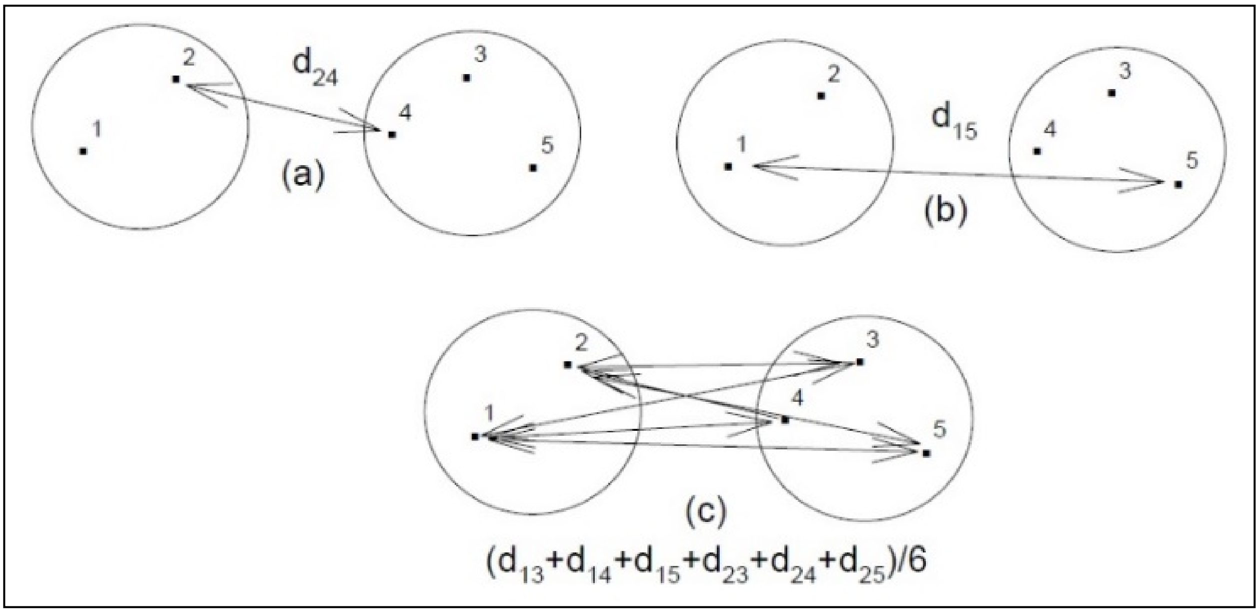

Figure 1. Grouping schemes according to the methods of connection: Single (a), Complete (b) and Average (c) (Johnson and Wichern, 2007)

the group with the shortest distance to the element. The pseudocode observing runs all objects and all groups doing permutations between them when needed (Algorithm 2).

Algorithm 2: Grouping based on k-means method

```
Input: Data $A_{n_x p}$ with $n$ measured objects in $p$ variables
Input: Number of groups $k$
Output: Matrix of distances $D_{n_x n}$ among the objects considering all the variables
Creates k groups randomly
While i = 1 → nthere is change in groups then
    Uses the centroid calculated to sort objects
    For $i = 1 \rightarrow k$ then
    Determines new centroid
    End
End
```

METHODOLOGY

This research is qualitative-quantitative (Appolinario, 2016), because it analyzes the characteristics of the participants of the sample in order to classify them into subgroups according to their technological abilities. For this, we use quantitative data generated from answers of survey applied with the respondents.

In addition, it has characteristics of the exploratory research, according to the classification of the research based on its objectives (Gil, 2002), because it intends to know better the problem and make it better known in the scientific community, including explaining the solution of the problem of research.

As for the technical procedures used (Gil, 2002), in the beginning, this research uses bibliographical and survey research. Bibliographic research is necessary for a better understanding of theoretical aspects of research, while survey research is necessary to collect data from the samples in order to be analyzed by the statistical methods chosen for the research, in order to group the research participants, in accordance with his skills with technology.

The research, then, sought to know the skills of teachers with the use of technology. The sample consisted of 71 teachers from educational institutions in the Brazilian states of Amazonas, Alagoas, Bahia, Ceará and Pernambuco and followed the type of survey surveys based on lists of the random sample model of Von Baur and Florian (2009). Thus e-mail lists of members of educational institutions were used to send invitations to members to respond to the online questionnaire about their ICT skills, attitudes and frequency of use.

Online survey research is a trend, because of the following advantages: low cost, time (speed of delivery to the target public), ease of use, absence of spatial constraints and good response rate (Flick, 2013).

The data collection instrument used was the online questionnaire with objective items about the level of knowledge and frequency of use of technologies and tools used as a didactic resource or as a learning support,

Table 1. Items researched among the teaching community (Freire et al., 2011; Martins et al., 2010; Oliveira et al., 2010)

Survey with Ability in Technology and Education Support Tools			
1	Computer	**16**	Printed or online maps
2	Netbook/notebook/ultrabooks	**17**	Learning Objects (LO)
3	Smartphones	**18**	Online translators
4	Tablet	**19**	Text editor
5	Internet	**20**	Online text editor
6	CD	**21**	Spreadsheet Editor
7	DVD	**22**	Online spreadsheet editor
8	Blu-ray	**23**	Slideshow Editor
9	Home theater	**24**	Online slideshow editor
10	TV	**25**	Image editor
11	Datashow	**26**	Online Image Editor
12	Search Engine	**27**	Online Messengers
13	Video repositories	**28**	Social networks
14	Scratch	**29**	Software Development Tool
15	Blog	**30**	Online messengers for smartphones

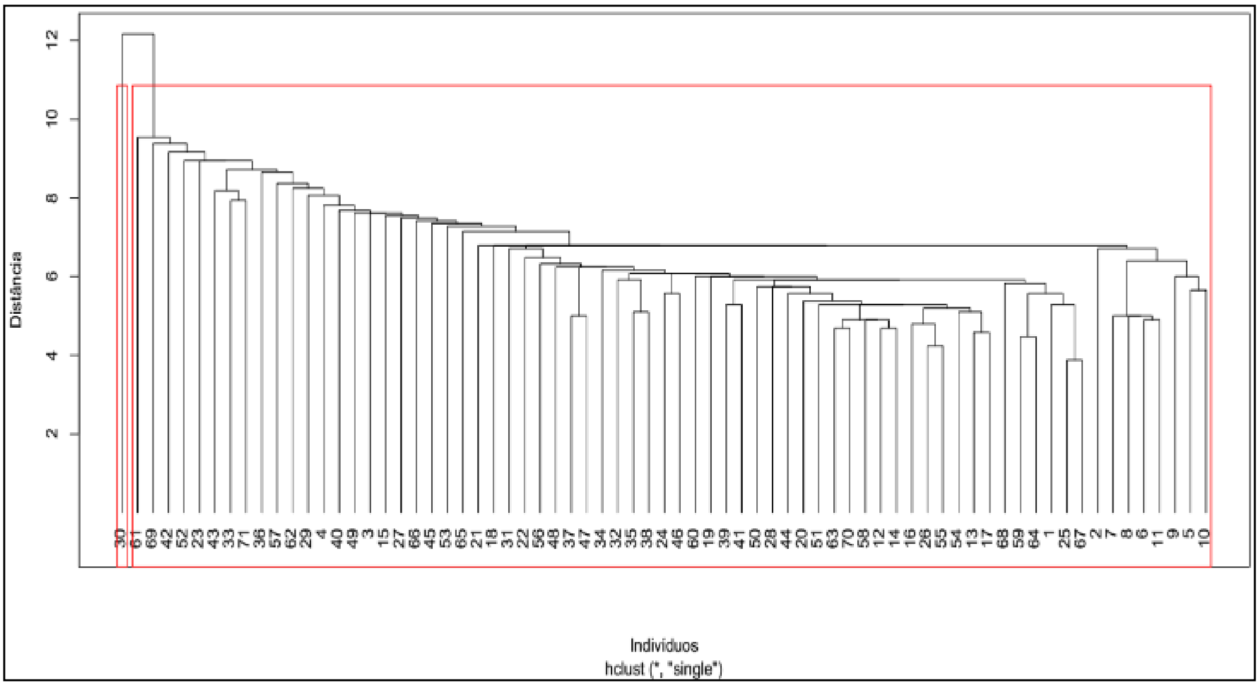

Figure 2. Single with LAT on the left and AT on the right

according to adaptations based on surveys of profile of teachers of (Freire et al., 2011), (Martins et al., 2010) and (Oliveira et al., 2010) and (Oliveira et al., 2016) (**Table 1**).

Frequencies of use followed the Likert scale (Marconi and Lakatos, 2002) with the following options: daily, weekly, monthly, semi-annual, annual, never used or unknown meaning. Participants received the form link in their email boxes as they were sent from institutional mailing lists. The first part of the survey consisted of a free and clarified term that presented the objectives of the research, the researchers involved, the participants 'rights, mainly, the guarantee of the exclusive use of the data for research purposes and with maximum confidentiality of identities of participants. Thus, the participant could accept or not participate in the research from the response to the survey item that inquired about such a situation. This survey remained online for about sixty days.

After this period the data were tabulated and formatted in spreadsheet editor so that the data were processed through software R, using the algorithm based on the grouping technique to define and create the subgroups. For this, the *skyeans*, *pvclust* and *cluster* packages were used to construct the graphs and execute the algorithms.

In the first moment, the profiles will be identified in relation to the abilities with use of technology of the sample from the use of the Cluster Analysis, also known as Cluster Analysis, Classification Analysis or Cluster Analysis. In this case, the main objective is to group the elements of the sample, according to the characteristics answered in the survey on the use of technology, as they present similarities among each other, because if we consider the total sample it will be possible to perceive heterogeneous characteristics among the sample individuals (Mingoti, 2005).

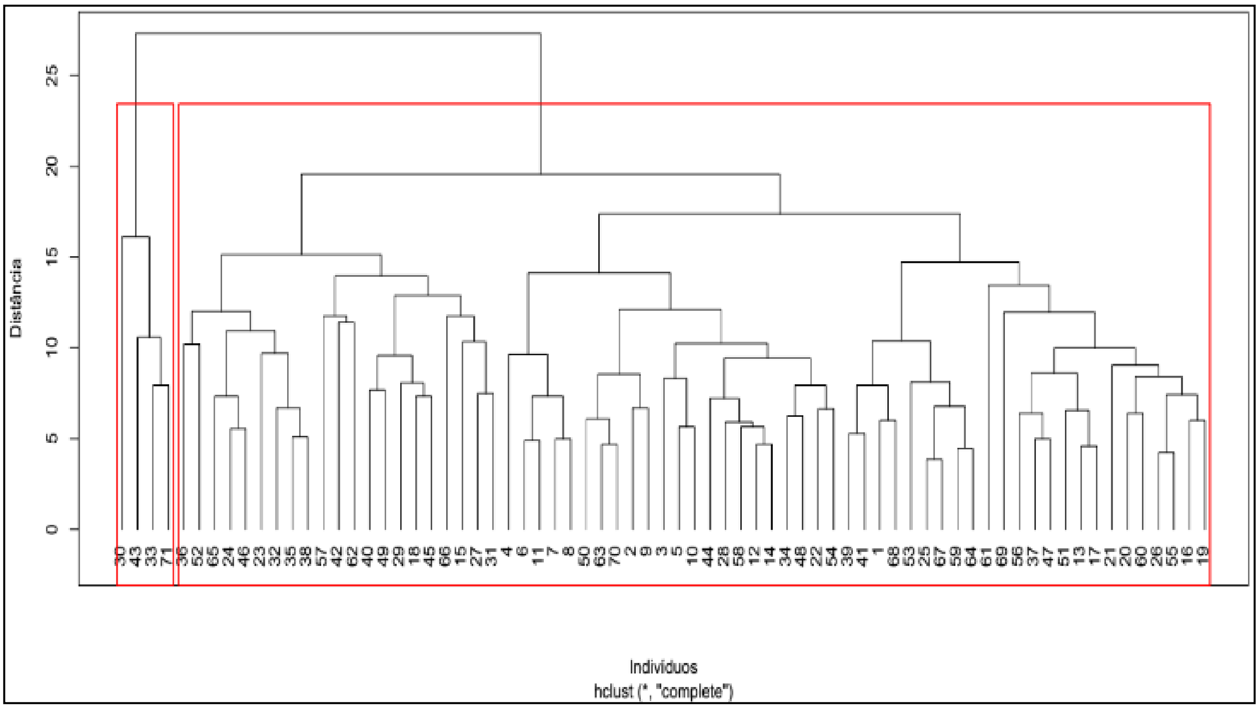

Figure 3. Complete with LAT on the left and AT on the right

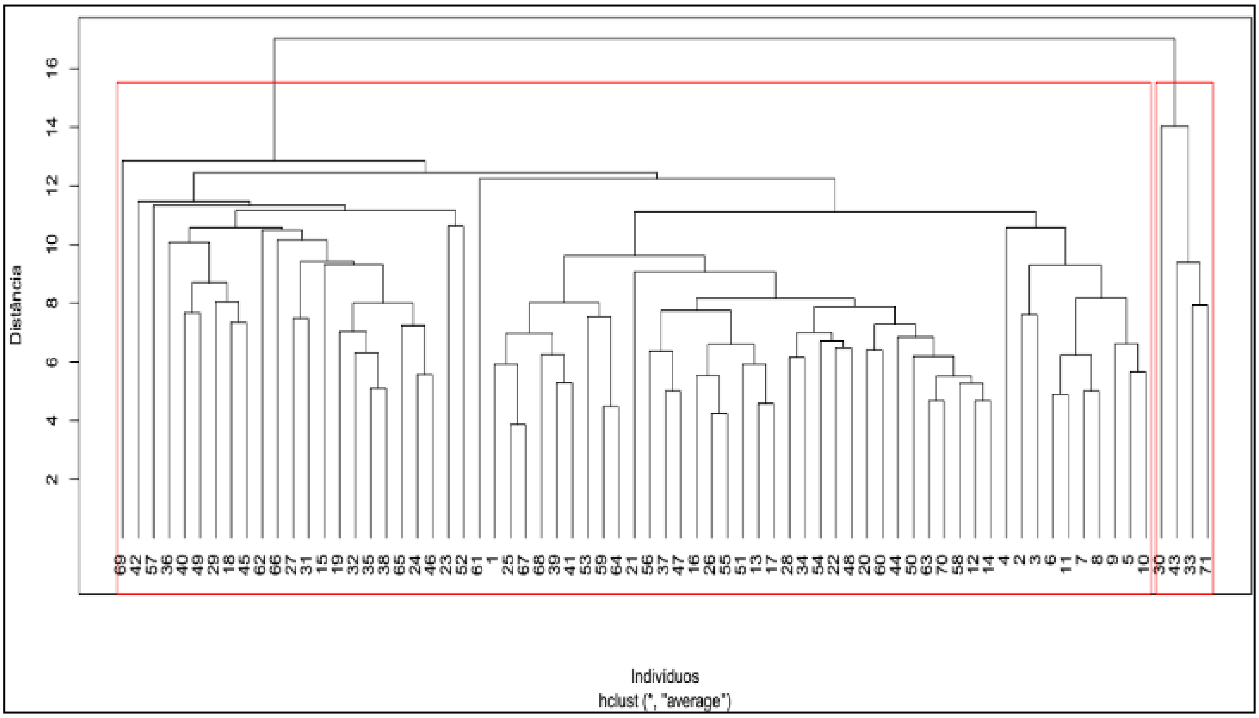

Figure 4. Average with AT on the left and LAT on the right

The representation of the data will be presented in the next section through graphs, specifically, dendrograms that brought together groups of users with ability in technology (AT) and low ability in technology (LAT).

RESULTS AND DISCUSSION

The groups were constituted from similar characteristics existing among the individuals of the sample, who presented them in the phase of data collection survey when answering the questionnaire on profile survey. The algorithms of the hierarchical method were applied in Single (**Figure 2**), Complete (**Figure 3**), Average (**Figure 4**) and Ward (**Figure 5**).

The **Figures 2, 3, 4** and **5** show dendograms of the hierarchical method with sample distribution among a group of users with skill with technology (AT) and with little ability with technology (LAT).

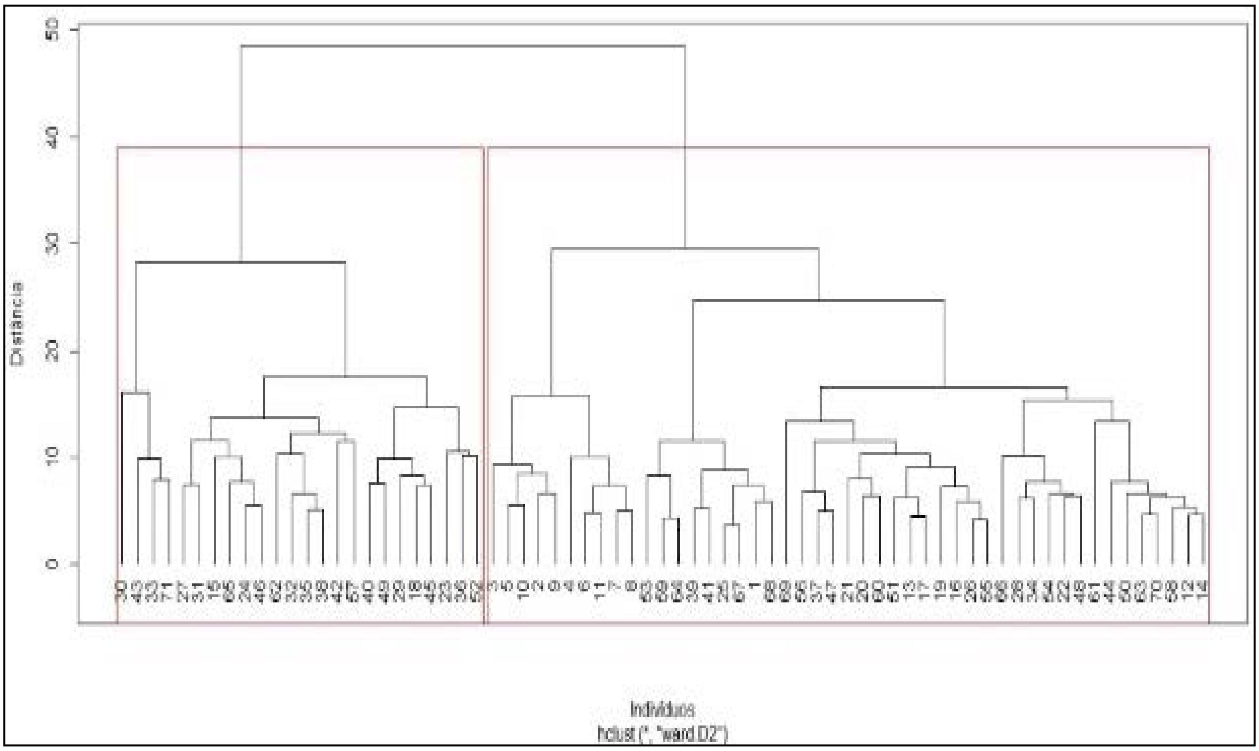

Figure 5. Ward with LAT on the left and AT on the righ

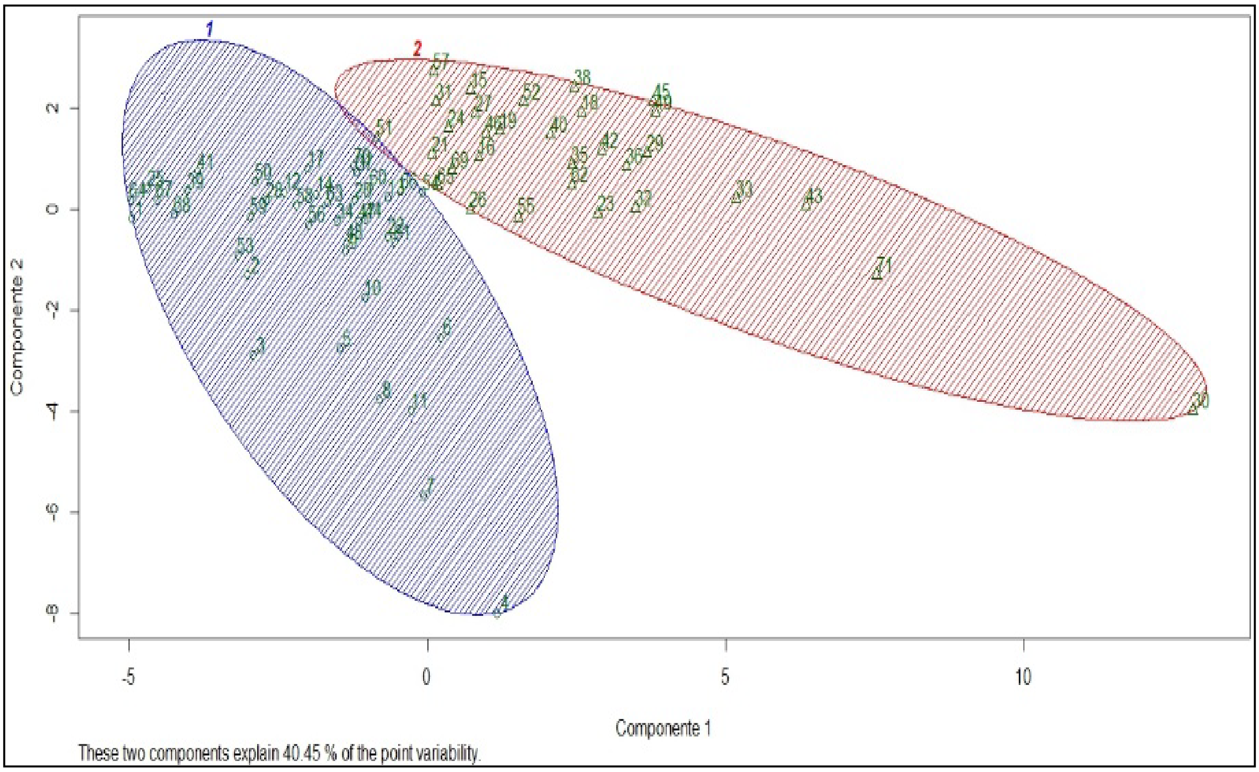

Figure 6. Representation of grouping of users with k-means based on main components

The four methods are based on Euclidean distance. Therefore, the more similar the responses of individuals, the smaller the distance between them. The analysis of the dendograms, generated from the data of the survey, allows to conclude that Ward grouping algorithm (Figure 5) can present a more coherent configuration of the individuals in each subgroup. This result was already expected, considering the robustness of the method by homogenizing the groups exhaustively in the iterations.

The algorithm of the non-hierarchical method (Figure 6) allows the visualization of the two subgroups in order to compare with Ward result (Figure 5). In both, the grouping is the same. The individuals, measured in the 30 variables, are placed in function of linear combinations. In Figure 6, the main components technique was used. In Figure 7, the discriminant function was used, providing a greater power of discrimination, as the name suggests.

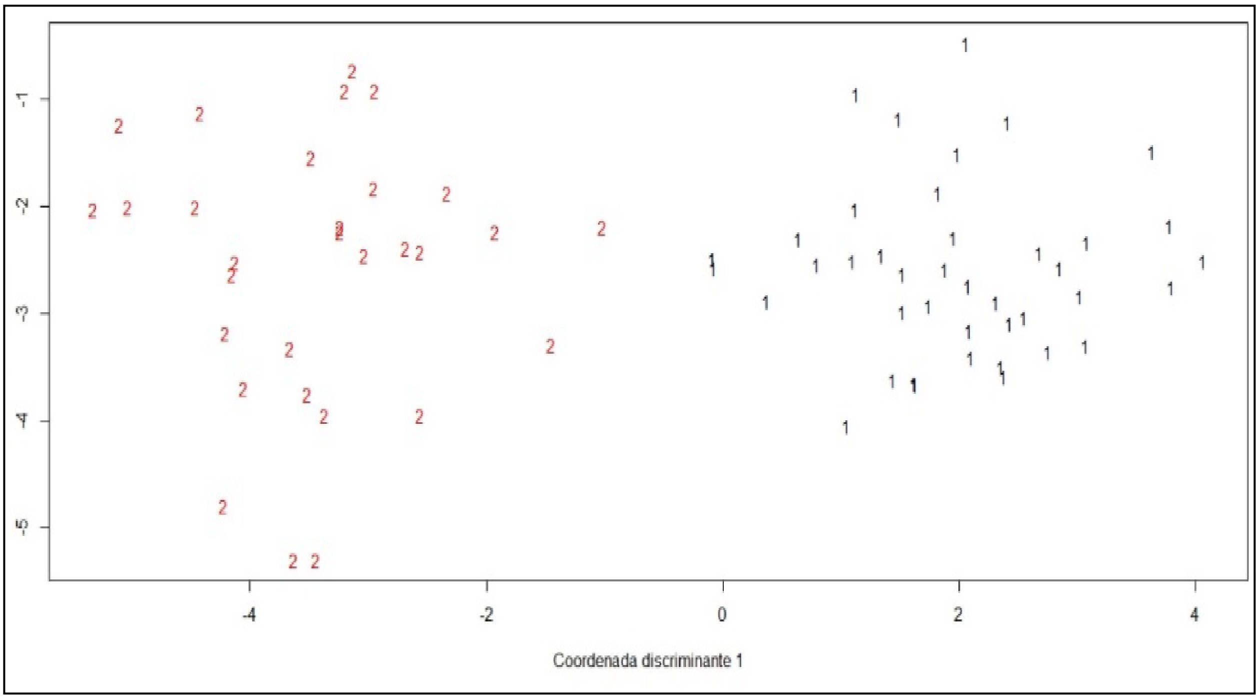

Figure 7. Representation of grouping of users with k-means based on discriminant coordinates

Table 2. Grouping according to Ward and K-means methods

	Ability with Technology (AT)		Low ability with Technology (LAT)	
Ward	1-14; 17; 20; 22; 25; 28; 34; 37; 39; 41; 44; 47; 48; 50; 51; 53; 54; 56; 58-61; 63; 64; 66; 67; 68; 70;	16; 19; 21; 26; 55; 69;	15; 18; 23; 24; 27; 29; 30; 31; 32; 33; 35; 36; 38; 40; 42; 43; 45; 46; 49; 52; 57; 62; 65; 71;	
K-means				16; 19; 21; 26; 55; 69;

For more details, see (Johnson and Wichern, 2007). In **Figures 6** and **7**, the plotted points are the elements rewritten in function of the respective linear combinations after the execution of the algorithm used by the method.

By grouping the survey participants using Ward and K-means, we can see the similarity of 92% of the users grouped in the profiles AT and LAT. There was an 8% divergence in the groupings performed by the two mentioned techniques, which occurred by the inclusion of 6 individuals in the group of users with low ability by the k-means technique, while Ward grouped the individuals in the profile AT (**Table 2**).

Therefore, when analyzing the profiles of the users in each group, we can see items 2, 12, 5, 4, 30, 13, 19, 28, 1, 11 of **Table 1** are the most relevant for defining the users in the AT group or LAT, while items 14, 8, 29, 26, 9 and 6 are the least relevant.

CONCLUSIONS

The use of these clustering methods is a convenient resource for research whose purpose is to segregate elements into subgroups impartially, free of subjective intervention. The grouping is performed through algorithms that have the Euclidean distance as execution criterion. Therefore, it is not known which variable has the greatest influence among those used, since the measure concentrates multivariate information into a single value.

The technique is purely descriptive, not allowing the application of hypothesis tests or the making of inferences. However, it is appropriate in scenarios with the purpose of classification in which the response pattern of the individuals belonging to each specific group is unknown, especially when using many variables. That is, it is not previously known the profile of an individual with technology ability for the subsequent association of an element with their profile.

After defining the groups, it is possible to identify the similar characteristics of the individuals that make up each one. In this way, it becomes possible to develop strategies to provide capabilities in specific characteristics to make individuals with little knowledge with technology can reach the group with skill with technology.

Thus, it is possible to abstract such sets of procedures from this research to other scientific applications in several areas of knowledge and also yearn for the representative selection of the sample, according to established research needs.

ACKNOWLEDGEMENT

The gratitude goes to Foundation for Research Support of Pernambuco (FACEPE), Federal Institute of Education, Science and Technology of Sertão Pernambucano (IF Sertão-PE) and Federal University of Pernambuco (UFPE), as well as to the professionals involved.

REFERENCES

Appolinario, F. (2016). *Metodologia da Ciência: Filosofia e prática da pesquisa*. 2. ed. São Paulo: Cengage Learning.

Barbosa, S. D. and Silva, B. S. (2010). *Interação Humano-Computador*. Rio de Janeiro: Elsevier.

Baxter, K., Courage, C. and Caine, K. (2015). *Understanding your users: a practical guide to user research methods*. 2nd. ed. San Francisco, CA: Morgan Kaufmann Publishers.

Bieniasz, A. and Majchrzak, A. (2011). Applying the Ward method in the analysis of financial situation of commercial banks. *e-Finanse: Financial Internet Quarterly*, 7(3), 1–12. Available at: https://www.econstor.eu/handle/10419/66756 (Accessed 1 March 2017).

Courage, C. and Bazter, K. (2005). *Understanding your users: a practical guide to user requirements, methods, tools, and techniques*. San Francisco, CA: Morgan Kaufmann Publishers.

Everitt, B. and Hothorn, T. (2011). Cluster analysis. In: *An Introduction to Applied Multivariate Analysis with R*. New York, NY: Springer. https://doi.org/10.1002/9780470977811

Flick, U. (2013). *Uma introdução à pesquisa qualitativa – um guia para iniciantes*. Penso: Porto Alegre.

Freire, R. S., David, P. B. and Oliveira, F. K. D. (2017). Dialogicidade na Formação Online de Professores de Matemática. Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE), 1(1). Available at: http://www.br-ie.org/pub/index.php/sbie/article/view/1658 (Accessed 1 March 2017)

Gil, A. C. (2002). *Como elaborar projetos de pesquisa*. São Paulo: Editora Atlas.

Hackos, J. T. and Redish, J. C. (1998). *User and task analysis for interface design*. New York, NY: John Wiley & Sons.

Hardle, W. and Simar, L. (2007). *Applied multivariate statistical analysis*. Berlin: Springer.

Hartigan J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, 28(1), 100. https://doi.org/10.2307/2346830

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. NJ: Prentice Hall.

Lee, G. (1993). *Object-Oriented GUI Application Development*. NJ: Prentice Hall.

Marconi, M. A. and Lakatos, E. M. (2002). *Técnicas de pesquisa: planejamento e execução de pesquisas, amostragens e técnicas de pesquisas, elaboração, análise e interpretação de dados*, 5th ed. São Paulo: Atlas.

Martins, C. A. et al. (2010). Dinâmica Pedagógica em Educação a Distância: Caracterizando Aspectos Metodológicos de uma Disciplina de Graduação da UAB. In: *Workshop Brasileiro De Informatica Na Educacao*. Porto Alegre: SBC.

Mingoti, S. A. (2005). *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Belo Horizonte: Editora UFMG.

Oliveira, F. K. D et al. (2017). Grouping methods for identification of profiles of teachers with expertise in educational technology. In: (Álvaro Rocha, Ed.) *12th Iberian Conference on Information Systems and Technologies (CISTI)*, Lisboa. https://doi.org/10.23919/CISTI.2017.7975940

Oliveira, F. K. D., Oliveira, M. B. D. and Gomes, A. S. (2016). Métodos hierárquicos e não-hierárquicos de agrupamento para classificação de indivíduos. In: F.K.D. Oliveira and K.F. Abreu (Eds.). *Métodos e pesquisas em Educação*, 1. Ed (pp. 39–62). Editora Kiron: Brasília.

Oliveira, F. K. D., Santana, J. R. and Pontes, M. G. O. (2010). O vídeo como ferramenta educacional a partir de múltiplas plataformas. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, 1(1). Available at: http://www.br-ie.org/pub/index.php/sbie/article/view/1493 (Accessed 1 March 2017).

Oliveira Netto, A. A. (2010). *IHC e a Engenharia Pedagógica: interação humano computador*. Florianópolis: Visualbooks.

Peffer, M. E. and Renken, M. (2015). Science Classroom Inquiry (SCI) Simulations for Generating Group-Level Learner Profiles. In: *Conference in Computer Supported Collaborative Learning (CSCL)*. Available at: https://www.researchgate.net/profile/Maggie_Renken/publication/279291336_Science_Classroom_Inquiry_SCI_Simulations_for_Generating_Group-Level_Learner_Profiles/links/5665ef9708ae15e74634c22d.pdf (Accessed 1 March 2017)

Rencber, A. C. (2002). *Methods of Multivariate Analysis*, 2nd ed. New York, NY: Wiley. https://doi.org/10.1002/0471271357

Ribas, A. D. et al. (2012). Similarity clustering for data fusion in Wireless Sensor Networks using k-means. In: The 2012 International Joint Conference on Neural Networks (IJCNN), IEEE. https://doi.org/10.1109/IJCNN.2012.6252430

Rogers, Y., Sharp, H. and Preece, J. (2013). *Design de interação: além da interação humano-computador*, 3rd. ed. Porto Alegre: Bookman.

Von Baur, N. and Florian, M. J. (2009). *Stichprobenprobleme bei Online-Umfragen. In: Sozialforschung im Internet* (pp. 109–128). VS Verlag für Sozialwissenschaften: Wiesbaden. https://doi.org/10.1007/978-3-531-91791-7_7