NaSTech JOURNAL OF NATURAL SCIENCES AND TECHNOLOGIES
2025, 4(1), pp. 396-399, DOI: 10.5281/zenodo.16595406

USING CLUSTER ANALYSIS METHODS TO CLASSIFY USERS OF THE EDUCATIONAL SPACE

Vusala Xudashirin MURADOVA

Faculty of Agriculture and Engineering, Department of Technology and Technical Sciences, Lankaran State University, Lankaran, Azerbaijan

ABSTRACT

Preliminary analysis allows, at the early stages of developing an educational space model, to determine the feasibility of designing information technology (including in the Client Server architecture) or distributed databases based on an analysis of the commonality of the educational space of a given set of users. The current article aims to the proposed method of community analysis is based on the assessment of the similarity function of users of the educational space.

Keywords: methods of cluster analysis, educational space, client server architecture, method of sequential acquisition`

1. INTRODUCTION

Preliminary analysis allows, at the early stages of IEP development, to determine the feasibility of designing information technology (including in the "Client Server" architecture) or distributed databases based on an analysis of the commonality of the educational space of a given set of users.

2. THE PROPOSED METHOD OF COMMUNITY ANALYSIS IS BASED ON THE ESTIMATION OF THE SIMILARITY FUNCTION OF USERS OF THE EDUCATIONAL SPACE.

Let be $U=\{u_k\}, k=1,K_o$ the set of users of the designed educational environment, which are given at the previous stage. Let be $D_k=\{d_{ki}\}, i=1,L_k^0$ the set of information elements describing the educational space of the k-th user and be the number of information elements in the set. The set D is the complete set of information elements of the considered set of users. It is formed by combining the information elements of the sets D_k.

After merging, the set is ordered so that identical information elements belonging to different user information sets are replaced by one, i.e.

$$D=D_1\cup D_2-D_{1,2}\cup D_3-(D_{1,3}\cup D_{2,3})\cup\ldots\cup D_{k0}-\bigcup_{i=1}^{k_0-1}D_{ik_0} \quad (1)$$

The relation of each user to the complete set of information elements D is given as a binary vector containing single elements in the position corresponding to an element from the information set of this user and zeros in all other positions. The set of such vectors for all users constitutes a binary matrix $B_k=\|b_{kj}\|$, indexed along the axes by a set of users $U=\{u_k\}$ and a complete set of information elements $D=\{d_j\}, j=1,n_0$ where is n_0 - the number of elements in the complete set. The elements of the matrix $B_k=\|b_{kj}\|$, take unit values if the information element d_j belongs to the information set of users U_k, otherwise $bk_j=0$.

At the first stage of the preliminary analysis stage, users are classified into two classes according to the degree of commonality of their educational space, and the set of users for whom individual design is appropriate is determined. The appropriateness is determined by the presence of a given degree of commonality between the educational spaces considered by the set of users.

Let be $D^0 = \{d_i\}, i = 1, N, N = \sum_k L_k^0$ the combined set of information elements obtained by combining the sets Dk , and which contains repeated information elements belonging to different sets Dk , i.e. $D^0 = \bigcup_{k=1}^{k_0} D_k$

Let us D_k^y define the set as a subset of the combined information set, which is D^0 defined as follows:

$D_k^y = D^0 - D_k, k \in 1, K_0$ (2)

The degree of commonality of the educational space of users is determined by the method of sequentially obtaining and analyzing pairwise intersections of the information sets of users with the corresponding sets D_k^y.

Let be the set of elements of the information set of an individual user Dk and the subset of users. If the D_k^p subset is not empty and the intersection power satisfies the given value, then the educational space of the kth user has a sufficient degree of commonality with the educational space of other users, which allows us to consider it as an individual user, distributed according to the geographical distribution of users in the computing environment (CO). To obtain a quantitative characteristic of the degree of commonality (power of intersection D_k^y) of the educational space of the kth user (which is given by the set), we will use the concept of a similarity measure, which is used in the theory of automatic classification.

Let Dk and D_k^y . Then the similarity measure is the mapping of the intersection of the sets $D_k^p = D_k \bigcap D_k^y$ onto some set of real numbers S_k , expressed by a non-negative real function, satisfying the condition $0 \le S_k \le 1$.

In the theory of automatic classification, a number of functions are used to calculate similarity measures between objects described in the form of binary vectors (functions of Russell and Rao, Sokal and Michener, Jacquard, Chuprov, etc., table 1).

Table 1. Functions for calculating similarity measures between objects

Code	Standard form of similarity function	Type of function in matrix model
S_1	$\frac{p_{11}}{p_{11}+p_{10}+p_{01}}$	$\frac{p_{11}}{n_0}$
S_2	$\frac{p_{11}+p_{00}}{p}$	$\frac{p_{11}}{n_0}$
S_3	$\frac{p_{11}}{p}$	$\frac{p_{11}}{n_0}$
S_4	$\frac{2p_{11}}{2p_{11}+p_{10}+p_{01}}$	$\frac{2p_{11}}{p_{11}+n_0}$
S_5	$\frac{2(p_{11}+p_{01})}{p_{11}+p_{10}+p_{00}}$	$\frac{2(p_{11}+p_{01})}{p_{11}+n_0}$
S_6	$\frac{p_{11}}{p_{11}+2(p_{10}+p_{01})}$	$\frac{p_{11}}{n_0+p_{10}+p_{01}}$
S_7	$\frac{p_{11}+p_{00}}{p+p_{10}+p_{01}}$	$\frac{p_{11}}{n_0+p_{10}+p_{01}}$

Most similarity functions include a quantity p_{00} , which determines the number of elements simultaneously absent from both considered sets. Since, by definition $D^0 = \bigcup_k D_k$ and $D_k^y = D^0 - D_k$, the quantity p_{00} always takes on a zero value. Thus, the use of a number of similarity functions (S_1, S_2, S_3) leads to degenerate estimates that do not ensure the adequacy of the comparison. Analysis $S_4 \ldots S_7$ of similarity functions shows that they ensure comparability of results. Among them, function S_6 (in the matrix model, function S_7 also corresponds to it) most accurately reflects the degree of commonality between two sets, since it allows taking into account

common (p_{11}) and specific (p_{10} and p_{01}) information elements in both sets. Therefore, it is proposed to calculate the similarity measure using the expression S_6, where $n_0 = p_{11} + p_{10} + p_{01}$ is the total number of different information elements in the considered sets, equal to the number of elements in the complete set $D = \{d_i\}$; (p_{11}) is the number of common elements in the information sets D_k and . The value of p11 is proposed to be calculated based on the characteristics of the matrix $B_k = \|b_{kj}\|$, as follows:

$$p_{11} = \sum_{j=1}^{n_0} Z_j \quad (3)$$

where Zj=1 if , $\exists : d_j \in D_k$ and $\sum_{k=1}^{k_0} b_{kj} \geq 1; z_j = 0$ – otherwise ;

p10 is the number of elements belonging to the set D_k but not in the set D_k^y. The value of p_{10} is calculated as

$$p_{10} = \sum_{j=1}^{n_0} X_j, \quad (4)$$

where Xj=1 if $\exists j : d_j \in D_k$ and $\sum_{k=1}^{k_0} b_{kj} = 1; X_j = 0$ otherwise;

p_{01}- the number of elements missing from the set D_k, but belonging to the set D_k^y, p_{01} is defined as

$$p_{01} = \sum_{j=1}^{n_0} Y_j, \quad (5)$$

where Yj=1 if $\exists : d_j \notin D_k$ and $\sum_{k=1}^{k_0} b_{kj} \geq 1; Y_j = 0$, otherwise.

Taking into account the notations (1–3), the similarity function S_6 takes the form:

$$S_k = \frac{\sum_j Z_j}{\sum_j Z_j + 2(\sum_j X_j + \sum_j Y_j)} \quad (6)$$

We assign a membership $U = \{u_k\}$ relation R to the set of users, which is determined by the value of the similarity measure from the set $S = \{S_k\} : U_k \in U_{РБД} \xleftrightarrow{R} S_k \geq S^*$, where $U_{РБД^*}$ the set of RBD users distributed over the OS nodes, S*- is the critical similarity measure.

The ratio R of users belonging to the class $U_{ЛБД}$ for which it is appropriate to design a set of individual (local) databases is defined as $U_k \in U_{ЛБД} \xleftrightarrow{R} S_k \geq S^*$.

For this group of users, the next stage of analysis involves structuring their educational spaces and building canonical structures of local databases.

The correlation $S_k \geq S^*$ is usually fulfilled in the case of a large number of geographically distributed users who solve information management and processing tasks that are similar in functional purpose and characteristics.

For a group of users $U_k \in U_{РБД}$, the characteristics of educational spaces are further analyzed in detail, external models, a generalized external model, and the canonical structure of the RBD are constructed.

Let us consider the use of the methods proposed in this section to solve a number of problems related to the use of a "Client-Server" network architecture.

In this regard, it is advisable to use the results of cluster analysis of information and classification of the educational space of users to determine the information composition of network databases (NBDs) located on the server.

For this purpose, a non-profit subset of information elements $D_{РБД} \subseteq D : D_{РБД} = D_k \bigcup D_{k'} \bigcup ... \bigcup D_k$ is defined for a subset of users $U_k \in U_{РБД}$, where $D_k, D_{k'}, ..., D_k$ is the set of information elements of users of the educational space $U_k \in U_{РБД}$.

Further, by varying the values of the critical similarity measure S*, groups of users (clusters) are determined,

which are distinguished based on the commonality of their educational spaces. For this purpose, the following analysis procedures are used:

Cluster analysis procedures similar to those considered above are performed on a set of users $U_k \in U_{РБД}$, but only for the magnitude of the similarity measure $S_0 < S^*$.

As a result of performing cluster analysis procedures, a set of working group users (WGU) $U_k^0 \in U_{СБД}$ $U_k^0 \in U_{СБД}$ and a set of local users are formed $U_k^0 \in U_{ЛБД}$.

The procedures discussed above are repeated until either further classification becomes impossible due to the technical capabilities of the selected LAN architecture (number of workgroups, network hubs, cabling, etc.), or for the remaining users, further reduction of the similarity measure becomes economically inexpedient, i.e. the degree of individuality (specificity) of the educational space is much higher than the degree of commonality and it is more profitable to duplicate individual similar information elements in local databases than to specifically allocate server (or workstation) capacity for this.

3. CONCLUSION

Thus, the analysis procedures considered above allow forming a hierarchical structure of minimally informationally interconnected user clusters, for each of which the information composition of network databases is subsequently determined.

REFERENCES

[1] Muradova V. Kh. Influence factors of decision-making systems on enterprise management PDMU-2022 xxxvii internationallconference problems of decision making under uncertainties november 23-25, 2022 Sheki-Lankaran, Republic of Azerbaijan November 23-25, 2022 P. 85 http://pdmu.univ.kiev.ua/PDMU_2022/PDMU_2022_Sheki.pdf

[2] Murad Omarov, Vusala Muradova BAYESIAN REGULARIZATION OF LEARNING Journal of Natural Sciences and Technologies ISBN:978-605- 73552-2-5 ICONAT-2022, Antalya, Turkey, 24-25 August, P. 80 https://journalofnastech.com/index.php/pub/issue/view/1

[3] Мурадова В.Х., Формалізовані моделі освітнього простору інформаційних вимог користувачів у системах дистанційного навчання 29-й Міжнародний молодіжний форум «Радіоелектроніка і молодь у XXI столітті». Зб. матеріалів форуму. Т. 7. Харків: ХНУРЕ. 2025. 291-293 с. УДК 004:37.018.43]:37.014.6