# The Correlation Of User's Interaction Activity On The Online Course Forum And User's Educational Achievement

**Nataliia D. Matrosova, Dmitry G. Shtennikov, Anna E. Karmanova**

**Abstract:** In this study, the authors decided to find out whether there is a linear correlation between the activity on the forum and the educational achievement. The data set was taken from one MOOC where students were trained. After data processing, a correlation scheme was constructed for pairwise comparison of variables.

**Index Terms:** forum activity, achievement of users, correlation, MOOC, machine learning, laboratory work, online course.

## 1 INTRODUCTION

User interaction with the information system implies interaction not only with the database, content and interface, but also with technical specialists (if such a need arises) or with operators. Increasingly, chat bots are used for such interaction, which "react" according to a specific, embedded algorithm. Chatbots do their job well when there are typical questions asked by many users accessing the system. But there are also other information systems in which not all questions or appeals can be predicted in advance and put into an algorithm. One such area is e-learning. Any distance course implies the presence of feedback (interaction) of students with the teacher or students among themselves. In massive open online courses (MOOC), when the number of students reaches several thousand, it is difficult for a teacher to implement synchronous interaction and respond quickly to emerging issues. Thus, the forum is one of the most important elements of user interaction on a distance course. Messages on the forum can be accessible for a long time to all participants of the course, which allows you to familiarize yourself with the questions and answers of other users. The forum is an element of asynchronous communication, and is often used to discuss the most pressing issues and problems of the distance course. In forums, users can receive support and assistance not only from the technical specialist and teacher, but also from other students. As it known, only a small part of students completes the distance course (about 3% in 2017-2018) [1]. The authors of the study decided to find out whether there is a linear correlation between the activity of the student's interaction on the forum and his educational achievement on the online course.

## 2 LITERATURE REVIEW

In preparing this research the authors studied a number of articles published on similar topics. Adrienne Traxler, A. Gavrin, and Rebecca Lindelln [2] used network analysis to compare forum logs, and found that "…more central network positions are positively linked with course success …". Also authors studied correlation between forum participation (total threads + comments) and final grade. Another study [3] proposed "new assessment criteria to help solve problem [students who choose not to join to forums] from the perspective of the supernetwork".

## 3 MATERIAL AND METHODS

First of all, the authors suggest defining the terms and what will be meant by them. An online course (distance course) is "an organized, focused educational process built on the basis of pedagogical principles of e-learning, implemented on the basis of technical means of modern information technologies and representing a logically and structurally completed educational unit, methodically provided with a unique set of systematized electronic teaching and monitoring tools" [4].MOOC – is "a training course with mass interactive participation using e-learning technologies and open access via the Internet" [5]. For the purposes of this study, the authors will use the terms "distance course", "online course" and "MOOC" as synonyms. Cambridge Dictionary [6] define the forum as "a place on the internet where people can leave messages or discuss particular subjects with other people at the same time". By user interaction activity is meant the number of messages that the user has left on the forum and other parameters (see below). The user's academic performance on the course consists of his grades for intermediate, mid-term and final certification. The correlation coefficient or pair correlation coefficient is "a measure of the linear dependence of two random variables." [7] This coefficient is denoted by r and can be [-1;1]. The closer the modulus of the correlation coefficient to unity, the stronger is the relationship between the measured values. The lack of communication is characterized by a correlation coefficient equal to 0 or a value close to it. The correlation dependence is not absolutely complete and accurate. It includes a plurality of causes and effects of various orders. [8] The authors used a dataset obtained as a result of teaching one stream (fall 2018) of students on the online course from the national open educational platform of the Russian Federation. Only those users who left at least one forum post were left in the dataset. The online course was a MOOC lasting ten weeks with information material presented in the format of video lectures. All information was made in a single style, which eliminates


- *Nataliia D. Matrosova, Faculty of Software Engineering and Computer Systems, ITMO University, Saint-Petersburg, Russian Federation, n.d.matrosova@gmail.com*
- *Dmitry G. Shtennikov, Faculty of Software Engineering and Computer Systems, ITMO University, Saint-Petersburg, Russian Federation, dshtennikov@itmo.ru*
- *Anna E. Karmanova, Graduate School of Service and Trade, Institute of Industrial Management, Economics and Trade, Peter the Great St.Petersburg Polytechnic University, Saint-Petersburg, Russian Federation, aekarmanova@bk.ru*

the influence of design on the perception of any topic. As an intermediate certification, after each week of training, users performed laboratory works in a virtual laboratory or test. Thus, the course had 11 laboratory works, 3 tests, one midterm test (after 5 weeks) and one final test (at the end of the course). The forums on the course allowed either to respond to pre-created, fixed topics or to create your own topics. Thus a user message could only be of one type: 1) main - the user independently created a new message thread. Such a branch could be of two different types (the user himself chose the type of message branch): a) discussion; b) question; 2) answer - the user replied in the thread created by the authors / technical assistants of the course or other users; 3) comment - the user responded to the user's response, thereby continuing the discussion. On the forums course were presented 651 posts, along with answers from the teacher / course author and technical assistants. Messages from the teacher / course author and technical assistants were used only to calculate additional parameters; such messages were deleted from the final selection. Since the course had a tight time frame, after which the laboratory work was not counted, additional variables were introduced: dl_week1, dl_week2, ..., dl_week10 - the end date of the intermediate certification for each week.

**The activity of user interaction in the forums is determined by the following indicator:**

1. count_main, count_answer, count_comment - the number of different types of messages within each week;
2. avg_count_words - the average length of messages in words within each week;
3. avg_count_symb - average length of messages within each week;
4. count_answered_tt - the number of messages of user received by the author of the course / technical assistants within each week;
5. count_answered - the number of messages / replies from users (for messages of the type “main” and “response”: for messages of the type “main” all messages of the type “reply” and “comment” are added up) within each week;
6. weekday - the most active day of the week within each week;
7. hourday - the most active hour interval (UTC + 02: 30: 28) within each week.

**The course has the following elements for assessing student knowledge:**

1. week 1 (deadline 15.10.2018): laboratory work 1 - 1 point, laboratory work 2 - 2 points, laboratory work 3 - 2 points;
2. week 2 (deadline 22.10.2018): laboratory work 4 – 2 points, laboratory work 5 – 2 points, laboratory work 6 – 1 points;
3. week 3 (deadline 29.10.2018): laboratory work 7 – 5 points;
4. week 4 (deadline 05.11.2018): laboratory work 8 – 3 points, laboratory work 9 – 2 points;
5. week 5 (deadline 12.11.2018): midterm test – 20 points;
6. week 6 (deadline 19.11.2018): laboratory work 10 – 5 points;
7. week 7 (deadline 26.11.2018): test 1 – 10 points;
8. week 8 (deadline 03.12.2018): laboratory work 11 – 5 points;
9. week 9 (deadline 10.12.2018): - test 2 – 5 points; test 3 – 5 points;
10. week 10 (deadline 17.12.2018): - final test – 30 points. For each element, the author used their own separate variable. Thus user’s educational achievement was determined by 16 variables. The dataset uses an estimate expressed as a percentage (100% for a correctly completed task).

Authors analyzed dataset and detected correlation using the Python programming language (Jupyter Notebook version 5.7.8) and additional libraries: pandas, datetime, matpotlob, seaborn, sklearn etc.

## 4 RESULT AND DISCUSSION

After grouping all the entries by ID, there were 115 users who left at least one message on the forums. As I.I. Eliseeva, M.M. Yuzbashev mentions in [10], “it is believed that the number of observations should be at least 5-6 times the number of factors (there is also a recommendation to use a proportion of at least 10 times the number of factors)”. The study uses 27 parameters (28th - ID), respectively, according to the recommendation, there should be at least 175-270 observations. There are fewer observations in the selected dataset, which may distort the results. Figure 1 presents the example of the dataset with variables to analyzing of user’s interaction activity on the online course forum.

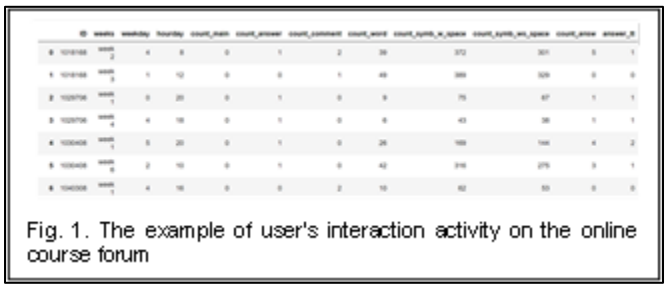

Fig. 1. The example of user's interaction activity on the online course forum

To connect the two datasets (with user activity on the forum and educational achievement), the authors used the pd.df.merge command, followed by pd.df.dropna and pd.df.fillna to clear the data. Figure 2 shows the graphs of the distribution of variables. It can be notice that variables aren’t normally distributed. However, the authors should notice that almost all of the laboratory works had only two variation of the marks – 0% or 100%. Then variables were normalized with command StandardScaler, that is the standard score of a sample x is calculated as (see Eq.1):

$$z = (x - u) / s \quad (1)$$

where u is the mean of the training samples or zero if with_mean=False, and s is the standard deviation of the training samples or one if with_std=False. [9]

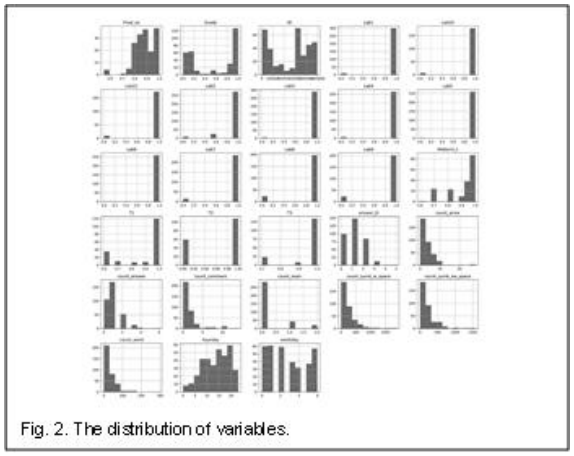

Fig. 2. The distribution of variables.

The Figure 3 illustrates the values of the correlation modulo and zeroed the diagonal elements, so that the cells with greater correlation were more noticeable.

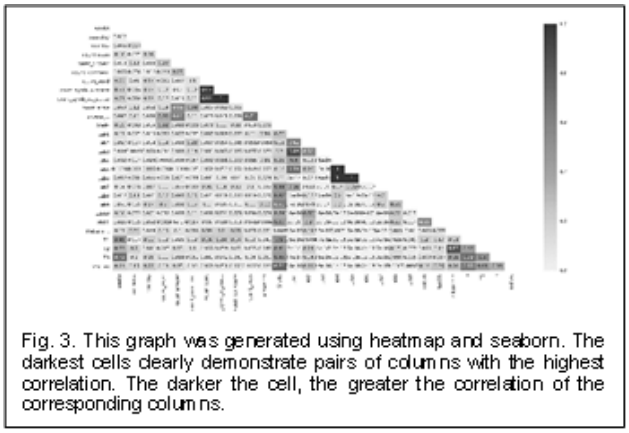

Fig. 3. This graph was generated using heatmap and seaborn. The darkest cells clearly demonstrate pairs of columns with the highest correlation. The darker the cell, the greater the correlation of the corresponding columns.

The correlation analysis showed that there is no linear strong relationship between the variables regarding the interaction on the forum and the variables regarding academic performance.

## 5 CONCLUSION

The study was conducted on a dataset obtained from one of the MOOC regarding the correlation of the user's interaction activity on the online course forum and the user's educational achievement. During the study, the data set was cleared of "noise" data. The remaining entries were analyzed for 28 variables. In conclusion, the authors would like to note that it is necessary to conduct additional research with a residual number of observations, as well as to consider the possibility of a nonlinear dependence between the variables. In conclusion, the authors would like to note that it is necessary to conduct additional research with a residual number of observations, as well as to consider the possibility of a nonlinear dependence between the variables.

## 6 REFERENCES


[1] D. Lederman, "Why MOOCs Didn't Work, in 3 Data Points," Inside Higher Ed, https://www.insidehighered.com/digital-learning/article/2019/01/16/study-offers-data-show-moocs-didnt-achieve-their-goals. 2019

[2] A. Traxler, A. Gavrin, and R. Lindell, "Networks Identify Productive Forum Discussions," Physical Review Physics Education Research, vol. 14, 020107, Published 10 September 2018 available at https://doi.org/10.1103/PhysRevPhysEducRes.14.020107

[3] C. He, P. Ma, L. Zhou and J. Wujiangw, "Is Participating in MOOC Forums Important for Students? A Data-driven Study from the Perspective of the Supernetwork," Journal of Data and Information Science, Volume 3: Issue 2, pp. 62-77, Published online: 22 Jun 2018, available at: https://doi.org/10.2478/jdis-2018-0009

[4] N.V. Grechushkina, "Online Course: Definition and Classification". Vysshee obrazovanie v Rossii = Higher Education in Russia. Vol. 27. No. 6, pp. 125-134. 2018 (In Russ., abstract in Eng.) Available at https://vovr.elpub.ru/jour/article/view/1403/1153

[5] O.V. Mereckov, "Do-it-yourself e-course development." Tutorial. 2019. Available at https://books.google.ru/books?id=5PSGDwAAQBAJ&lpg=PP1&hl=ru&pg=PP5#v=onepage&q&f=false

[6] Forum. Cambridge Dictionary. Available at https://dictionary.cambridge.org/dictionary/english/forum

[7] The correlation coefficient. The Math. Available at https://math.wikia.org/ru/wiki/%D0%9A%D0%BE%D1%8D%D1%84%D1%84%D0%B8%D1%86%D0%B8%D0%B5%D0%BD%D1%82_%D0%BA%D0%BE%D1%80%D1%80%D0%B5%D0%BB%D1%8F%D1%86%D0%B8%D0%B8

[8] M.M. Andreeva, V.R. Volkov, "Correlation analysis in sociological research," Vestnik Kazanskogo tekhnologicheskogo universiteta = Bulletin of Kazan Technological University. Vol. 7 pp, 2013. (In Russ ., abstract in Eng.). Available at https://cyberleninka.ru/article/n/korrelyatsionnyy-analiz-v-sotsiologicheskih-issledovaniyah/viewer

[9] sklearn.preprocessing.StandardScaler. scikit learn. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[10] I.I. Eliseeva, M.M. Yuzbashev "General Theory of Statistics: A Textbook." 4th edition, revised and supplemented. Finansy i Statistika = Finance and Statistics, Moskva, 2002,pp.228-230