Environment. Technology. Resources. Rezekne, Latvia
Proceedings of the 12th International Scientific and Practical Conference. Volume II, 218-221

The Personification of the User's Interface: Classification vs. Clusterization of Users of Online Courses

Nataliia D. Matrosova
Faculty of Software Engineering and Computer Systems
Saint-Petersburg National Research University of Information Technologies, Mechanics and Optics
Saint-Petersburg, Russian Federation
ndmatrosova@corp.ifmo.ru

Dmitry G. Shtennikov
Faculty of Software Engineering and Computer Systems
Saint-Petersburg National Research University of Information Technologies, Mechanics and Optics
Saint-Petersburg, Russian Federation
dshtennikov@corp.ifmo.ru

***Abstract*—Researchers compared the classification and the clusterization of users of online course for the personification of the users' information system interface. When interacting with control and information systems, users may manifest individual features, including implicit characteristics that may affect one's results within the system. At the same time due to information system building peculiarities one of the most comprehensive statistics can be collected via e-learning systems. When using a course, the user leaves a wide trail of activity that may contain different information depending on the learning environment structure. Online blended learning courses draw the researcher's attention to the impact of digital teaching models on students as well as its ability to adjust distant learning courses to individual students' needs and differences.**

Information personalization is a highly relevant content presentation at the most individual level. Therefore, the task of personalization is to show users information that meets their needs and interests. Personalization gives the opportunity to focus on points that have real value for users.

***Keywords*—machine learning, dataset, classification, clusterization, personification.**

I. Introduction

Firstly, it is important to clarify the term *user*. A user of information system is a specialist in the system's subject area, for whom the system was created to satisfy his informational needs.

The information system user interface consists of various elements. With a large amount of information, the user's in-system work efficiency decreases, due to attention diffusion. Through changes in the graphical part of the interface it is possible to implement recommendation modules. This change in interface is called user interface personification.

It is more efficient to carry out interface upgrades for each individual user than to personalize the interface for all users at once. However, with a large number of users there may be too many adapted options to store. The solution to this challenge is grouping users and conducting personification procedures for each group (cluster) separately.

The task of personification is to display information that meets the needs and interests of users. Personification allows you to focus users on important details. The simplest type of interface personification is a ranked list of items [1].

To achieve the best results in personification it is wise to use not only user explicit characteristics but also non-explicit ones, such as circadian rhythms for example. There are detailed studies of the effects of circadian rhythm on learning [2, 3].

Interface personification is based on algorithm adaptation, which describes rules by which interface changes depending on the user's actions. Among the characteristics of the adaptive algorithm, the following are emphasized: prediction of accuracy, predictability of adaptive behavior, and frequency of interface changes [4].

In this investigation the authors considered only the question of comparing classification and clustering for user interface personification on how to exert the influence of the circadian rhythm on learning. It did not address the question of whether all users from each identified class really require different user interface compared to users from another class.

A. Literature review

The possibilities of interface characterization for information systems, especially for training information systems, are discussed in detail in various studies.

The question regarding the characteristics of the

Print ISSN 1691-5402
Online ISSN 2256-070X
http://dx.doi.org/10.17770/etr2019vol2.4080

learner is discussed in detail in Hendrik Dranchsler and Paul A. Kirschner' article [5]. The authors suggest that learner characteristics can be personal (age, gender, maturation, language and etc.), academic (learning goals, prior knowledge, education type and level), social/ emotional (sociability, self-image, mood), cognitive (memory, mental procedures, intellectual skills) [5]. These characteristics are highly individual and vary for each learner.

Christoph Fröschl, Loc Nguyen, and Phing Do in the study [6] described an adaptive system based on the "description of learner's properties" (user model or learner model). In their work "the user model can contain information from two categories: domain specific information (reflects status and degree of knowledge and skills) and domain independent information (may include goals, interests, background and experience, personal traits and demographic information)" [6]. The authors offered to classify the user model into three kinds: stereotype, overlay, and plan models.

In the study of the influence of student characteristics on learning paths and strategies [7] the authors considered the following characteristics of students: prior knowledge, study level, gender, and intrinsic motivation. The results showed that students do indeed follow individual learning paths and some student characteristics are related to their learning paths (gender and prior knowledge did not have an effect, but intrinsic motivation had a stronger influence than prior knowledge) [7].

Using the classification and clustering of users to analyse the characteristics of users are considered Ronald G. Leppan, Johan F. van Niekerk, Reinhardt A. Botha in their study [8]. The authors suggested that online learning design should be informed by behavioural patterns. And learner characteristics are inferred using data analysis. The classification used as "predictive modelling to model something that cannot be directly observed by using readily available features as input" [8], and the clusterization - "structure discovery to find patterns in data that are not obvious" [8].

The authors suggest in further research to experiment with the personification of the user's interface to confirm this theoretical research.

II. Materials and methods

This research uses the dataset from one massive open online course (MOOC) from the national open education platform of the Russian Federation as source data.

The set contains data from one batch of students (spring 2018). The students were offered to study learning materials (lecture in video format), complete an after video mini-assessment, working in a virtual laboratory.

The data set contains over 900 000 logs from students and their activities with approximately 90 features. User classification and clusterization were compared using students' latent features and course success rates.

The equation for course user success can be written as Eq.1:

$$s = \sum_1^i \frac{g_i}{k} + \frac{\sum_1^l g_{ml}}{l} + \frac{\sum_1^p g_{fp}}{p} \quad (1)$$

Where s – course success, – after video mini-assessments grade i, –midterm grade, - final test grade, k, l, p – the number of examination passing tries (mini, midterm, and final accordingly).

A. Research data

After data preparation (null rows and Nan values deletion, choosing features), about 1450 users and 5 features were chosen (see Table 1).

It can be noted in Figure 1 that there is a weak correlation between features, but a strong one between weekday_video and weekday_lms, and also hour_video and hour_lms at this stage of the study.

TABLE I. choosing features from the dataset

Name features	Description
w e e k d a y _ video	the highest day of the week lection materials activity (lecture-activity weekday)
hour_video	the highest hour of a day lection materials activity (lecture-activity hour-day)
weekday_lms	the highest day of the week virtual laboratories activity (lms-activity weekday)
hour_lms	the highest hour of a day virtual laboratories activity (lms-activity hour-day)
grade	course student success rate
class	course user group (class) success (target variable for classification)

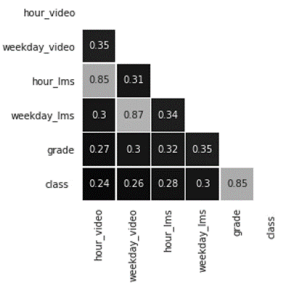

Fig. 1. Correlation between features

Figure 2 presents *the* distribution of features. It can be seen that most features do not have normal distribution.

B. Classification

Classification – the process of streamlining or distributing objects (observations) into classes in order to reflect relations between them [9].

Calculations were carried out for some classification models in this part of the investigation:

- classification and regression trees (CART) - solves classification and regression problems by building a decision tree;
- k-nearest neighbors algorithm (k-NN) – assigns the object to the class that is most common among its k-neighbors whose classes are already known;

- linear regression - estimates coefficients of the linear equation containing one or more independent variables, allowing better value prediction of the dependent variable;
- support vector machines (SVM) – has a special feature which is a continuous decrease in empirical classification error and an increase in the gap. The main idea of the method is a translation of initial vectors into space of higher dimension and search for separating hyperplane with the maximum a gap in that space. Fit time complexity is more than quadratic with the number of samples which makes it hard to scale to the dataset with more than a couple of 10000 samples. Therefore, the authors will not use this model further in this study;
- logistic regression - linear classifier construction method, which allows posterior probabilities evaluation of objects belonging to classes [10];
- Bayesian classification approach - is based on a theorem stating that if the distribution densities of each class are known, then the desired algorithm can be written in explicit analytical form. Moreover, this algorithm is optimal, as it has minimal error probability [10].

Table 2 was compiled as a result of the construction of all the above models. This table contains information about mean accuracy on given test data and labels. Note that linear regression contains a prediction coefficient of determination .

Also note that models were separately checked several times on test samples to avoid possible overfitting problems.

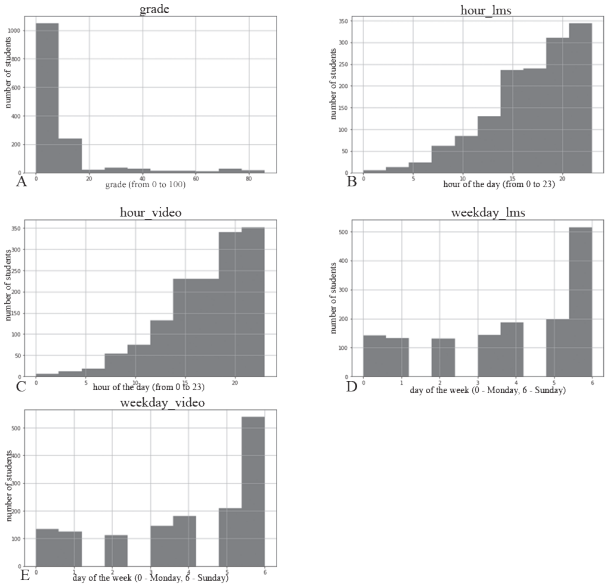

Fig. 2. Distribution of features: A – grade, B – hour_lms, C – hour_ video, D - weekday_lms, E - weekday_video

C. Clusterization

Clusterization (or cluster analysis) – the task of breaking up a set of objects into groups called clusters [11]. Clusterization involves the selection of compact, separate groups of objects characterized by internal homogeneity and external isolation.

In this investigation the authors used the best known method of clusterization - kMeans. To select an appropriate number of clusters, usually the number of clusters chosen from which (the sum of squares of distances from points to centroids of clusters to which they belong, see Eq.2) ceases to decrease sharply (Figure 3). In this example, the number of clusters is 4.

$$J(C) = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i \quad \mu_k\|^2 \rightarrow \min_C \tag{2}$$

TABLE II. MEAN ACCURACY ON THE GIVEN TEST DATA AND LABELS

Approach	Mean accuracy
classification and regression trees (CART)	0.91
k-nearest neighbors algorithm	0.86
linearregression	0.73
support vector machines (SVM)	0.83
logistic regression	0.66
Bayesian classification approach	0.84

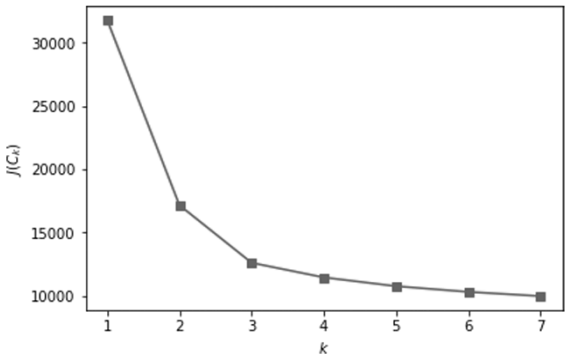

Fig. 3. The graph to define the number of clusters - the sum of the squares of distances from points to centroids of clusters to which they belong

```
[[0.46 -0.14 0.38 -0.88 -0.54]
 [-0.98 -0.08 -0.82 -0.18 -0.30]
 [0.04 0.00 0.01 0.05 1.51]
 [0.34 0.20 0.31 0.98 -0.50]]
```

Fig. 4. Coordinates of centroids

Coordinates of 4 cluster centers were received (Figure 4) as well as an additional column to data, containing information about cluster number to each user_id (Figure 5).

III. RESULTS AND DISCUSSION

Approaches to personification interfaces can be divided into two types: stereotypical and individual [12].

Stereotypical approach states that interfaces are

collected for several users' classes; the system classifies a user and provides one of these interfaces. Usually, several user models are created. For this situation and for predictive purposes (such as defining a user class) classification can be used. For example, the allocation of a user to a certain class based on one's activity and success.

As can be seen from Table 2, classification models give good predictive results. Especially, classification approaches and regression trees.

An individual approach personalizes the interface to improve the layout for a specific user (or a group of users with similar characteristics) based on behavioral data. Therefore, there are individual statistics on in-system actions for each group.

As authors stipulated above, classification and clusterization can address various tasks, as summarized in Table 3.

user_id	hour_video	weekday_video	hour_lms	weekday_lms	grade	cluster
300.0	13	3	13	3	0	1
429.0	15	6	16	6	13	2
525.0	20	6	20	6	2	2
715.0	23	5	23	5	6	2
853.0	14	6	14	6	4	1
1206.0	20	6	20	6	23	2
1874.0	21	6	21	6	82	3
3172.0	20	6	21	6	11	2

Fig. 5. New column in the dataset – cluster

TABLE III. CLASSIFICATION AND CLUSTERIZATION COMPARISON FOR PERSONIFICATION INTERFACE

Classification	Clusterization
Used for stereotypical personification.	Used for individual personification.
Needs prepared data.	Does not need prepared data.
The number of necessary changes for interface personification is reduced due to user grouping, which is defined as closest to the selected one.	Allows cluster centres to replace all users in the cluster, due to similar users clustering principles. Thus, the user database is formed.
Reduces the number of resources used and improves system performance.	Allows a reduction in search time for solutions and memory.
Possible decline in interface personification quality.	Loss of accuracy at cluster boundaries.

Conclusions

The authors studied classification and clusterization on a real data set and their effect on the personification a user's interface. It can be concluded that classification is better used in situations with prepared data, when users are already using a system and administrators can divide people into groups. On the other hand, clusterization is good at "cold start" situations.

For personification of user's interface classification may assist with stereotypical personification but clusterization – with individual personification.

Support for user personification (differentiation rules provision, interface adaptation, required information obtention) can distinguish an information system from a variety of similar competitive systems.

The authors additionally note that an experiment will be conducted to identify that the particular machine learning methods can be used for user interface personification.

References

[1] A. Schade "Customization vs. Personalization in the User Experience" // Nielsen Norman Group World Leaders in Research-Based User Experience.Jul. 10, 2016. [Online]. Available: https://www.nngroup.com/articles/customization-personalization/ [Accesses: Jan. 11, 2019].

[2] Mou X "What Role do Circadian Rhythms Play in Learning and Memory?" J Neurol Neurophysiol, 2016. 7:367. doi:10.4172/2155-9562.1000367. [Accesses: Jan. 15, 2019]

[3] P. Valdez, C. Ramírez and A. García "Circadian Rhythms in Cognitive Processes: Implications for School Learning". Mind, Brain, and Education, vol. 8, pp. 161-168, 2014 doi:10.1111/mbe.12056. [Accesses: Jan. 11, 2019]

[4] L. Findlater, Krzysztof Z. Gajos. "Design Space and Evaluation Challenges of Adaptive Graphical User Interfaces". AI Magazine. Vol. 30, No. 4. p. 68 – 73. 2009. [Accesses: Mar. 3, 2019].

[5] H. Drachsler, P.A. Kirschner "Learner Characteristics". // Encyclopedia of the Sciences of Learning. Chapter: Learning Characteristics. SpringerEditors: N. M. Seel. 2011. doi: 10.1007/978-1-4419-1428-6_347 [Accesses: Apr. 12, 2019].

[6] C. Fröschl, L. Nguyen, P. Do Learner Model in Adaptive Learning. The 2008 World Congress on Science, Engineering and Technology (WCSET2008), Volume 35, November 2008, Paris, France. 2008 [Accesses: Apr. 13, 2019].

[7] J.R. van Seters, M.A. Ossevoort, J. Tramper, M.J. Goedhart, The influence of student characteristics on the use of adaptive e-learning material, Computers & Education, Volume 58, Issue 3, 2012, Pages 942-952, ISSN 0360-1315, https://doi.org/10.1016/j.compedu.2011.11.002. [Accesses: Apr. 13, 2019]

[8] R.G. Leppan, J.F. van Niekerk, R.A. Botha "Process model for differentiated instruction using learning analytics". South African Computer Journal, 30 (2), pp. 17-43, 2018. doi: 10.18489/sacj.v30i2.481 [Accesses: Apr. 14, 2019]

[9] M.W.Sholom "Text mining. Predictive methods of analyzing unstructured information". M.W.Sholom, N.Indurkhya, T.Zhang, F.J.Damarau. — 2004. — 236 p. [Accesses: Jan. 22, 2019]

[10] MachineLearning.ru – [Online]. Available: http://www.machinelearning.ru/ [Accesses: Jan. 21, 2019].

[11] A.Chasovskih [Overview of data clustering algorithms] Obzor algoritmov klasterizacii dannyh"– Habr. Aug 11, 2010. [Online]. Available: https://habr.com/ru/post/101338/ [Accesses: Mar. 1, 2019].

[12] I.S. Morozov [Clustering methods of users for adaptationinterfaces] "Metody klasterizacii pol"zovatelej dlja adaptacii interfejsov", Bachelor thesis, Novosibirsk State University, Novosibirsk, Russian Federation, 2013. [Accesses: Feb. 27, 2019].