Additional Report - Detecting and Clustering Students by their Gamification Behavior with Badges

Jose A. Ruiperez-Valiente, Pedro J. Muñoz-Merino, Carlos Delgado Kloos

17 de mayo de 2016

Summary

This is an additional report to the article titled as "Detecting and Clustering Students by their Gamification Behavior with Badges: A Case Study in Engineering Education" and submitted for publication to the International Journal of Engineering Education. The report contains an exploratory analysis, correlation among the different metrics, analysis by plot of specific students and the clustering of students by badge indicators, as described in the paper.

Loading libraries

```
library(ggplot2)
library(reshape2)
library(dplyr)
library(Hmisc)
library(corrplot)
library(grid)
```

Preparing data

```
# Load data
allStudents <- read.csv("allStudents.csv", header = TRUE, sep = ";" , dec = ",")
differentStudents <- read.csv("differentStudents.csv", header = TRUE, sep = ";", dec = ",")

# Remove students who did not participated for at least 60 min from total
allStudents <- allStudents[allStudents$Total_Time >= 60,]
```

Exploratory analysis and descriptive statistics

```
# Metrics used within the analysis
metricsAnalysis <- c("User_id", "Course", "Cluster_TwoSteps", "Topic_Intention", "Repetitive_Intention"

metricsBadges <- c("Topic_Intention", "Repetitive_Intention", "Concentration", "Badges_Per_Time")

metricsOthers <- c("Cluster_TwoSteps", "Total_Time", "Proficient_Exercises", "Completed_Videos", "Exerci

# Structure of the dataframe used for the analysis
str(differentStudents[,metricsAnalysis])
```

```
## 'data.frame':    291 obs. of  13 variables:
##  $ User_id             : Factor w/ 289 levels "http://moodleid.khanacademy.org/uc3m/202",..: 2 5 11 1
##  $ Course              : Factor w/ 3 levels "2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Cluster_TwoSteps    : Factor w/ 3 levels "1","2","3": 1 1 2 3 1 2 2 1 1 2 ...
##  $ Topic_Intention     : num  66.7 50 0 0 28.6 ...
##  $ Repetitive_Intention: num  50 28.6 0 66.7 53.9 ...
##  $ Concentration       : num  95.6 72.2 0 0 59.3 ...
##  $ Badges_Per_Time     : num  3.53 2.02 1.63 2.38 3.04 0.84 0.59 5.64 2.33 1.24 ...
##  $ Total_Time          : int  122 668 261 67 616 231 95 392 218 338 ...
##  $ Proficient_Exercises: int  13 56 3 3 46 0 0 60 13 3 ...
##  $ Completed_Videos    : int  10 64 33 0 85 19 10 32 19 62 ...
##  $ Exercise_Abandon    : int  0 22 90 80 46 100 100 40 20 92 ...
##  $ Video_Abandon       : int  25 18 9 100 7 28 75 93 0 50 ...
##  $ Optional_Elements   : int  16 0 0 16 0 16 0 16 0 0 ...
```

```
# Summary of the badge indicators
summary(differentStudents[,metricsBadges])
```

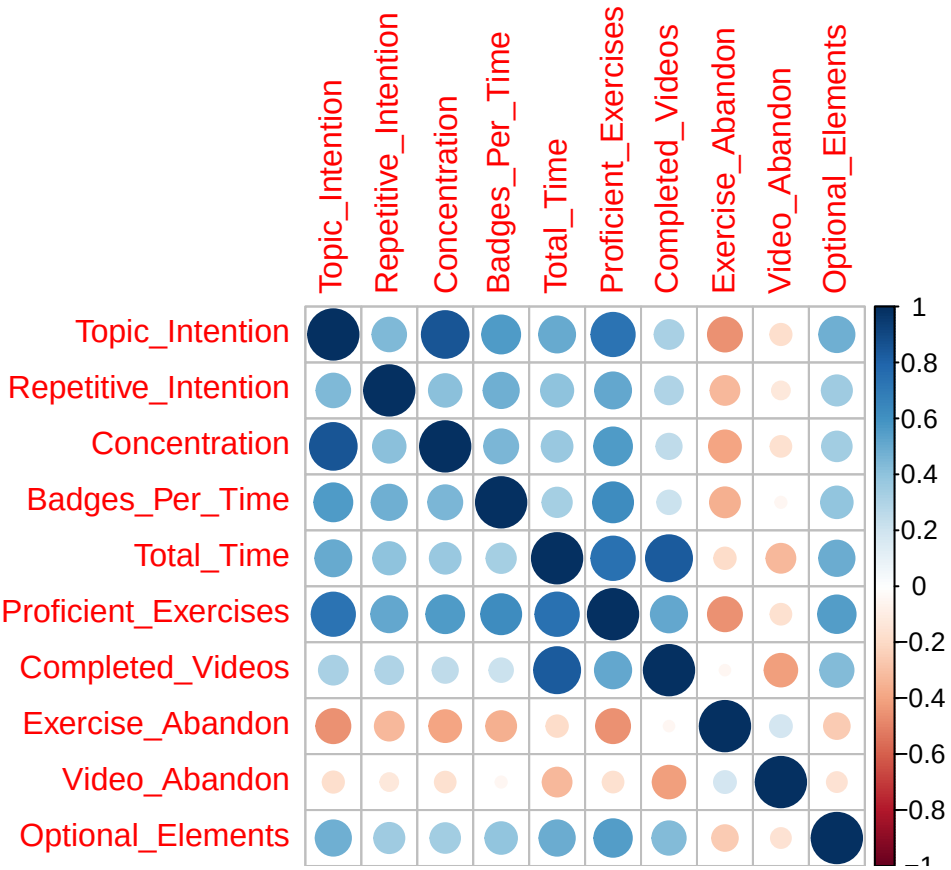
```
##  Topic_Intention  Repetitive_Intention Concentration    Badges_Per_Time
##  Min.   :  0.00   Min.   :  0.00       Min.   :  0.00   Min.   : 0.000
##  1st Qu.:  0.00   1st Qu.:  0.00       1st Qu.:  0.00   1st Qu.: 1.025
##  Median :  0.00   Median :  0.00       Median :  0.00   Median : 1.950
##  Mean   : 15.18   Mean   : 32.74       Mean   : 16.54   Mean   : 2.564
##  3rd Qu.: 21.11   3rd Qu.: 62.91       3rd Qu.: 20.61   3rd Qu.: 3.525
##  Max.   :100.00   Max.   :100.00       Max.   :100.00   Max.   :19.790
```

```
# Summary of the rest of the indicators used in the study
summary(differentStudents[,metricsOthers])
```

```
##  Cluster_TwoSteps   Total_Time     Proficient_Exercises Completed_Videos
##  1: 72            Min.   :   0.0   Min.   :  0.00       Min.   :  0.00
##  2:149            1st Qu.:  54.0   1st Qu.:  0.00       1st Qu.:  4.00
##  3: 70            Median : 177.0   Median :  5.00       Median : 18.00
##                   Mean   : 261.2   Mean   : 16.79       Mean   : 30.66
##                   3rd Qu.: 364.0   3rd Qu.: 26.00       3rd Qu.: 50.00
##                   Max.   :2458.0   Max.   :100.00       Max.   :100.00
##  Exercise_Abandon Video_Abandon   Optional_Elements
##  Min.   :  0.00   Min.   :  0.0   Min.   :  0.000
##  1st Qu.: 24.00   1st Qu.:  0.0   1st Qu.:  0.000
##  Median : 62.00   Median : 28.0   Median :  0.000
##  Mean   : 56.87   Mean   : 38.1   Mean   :  7.663
##  3rd Qu.:100.00   3rd Qu.: 67.5   3rd Qu.:  8.000
##  Max.   :100.00   Max.   :100.0   Max.   :100.000
```

Relationship between badge indicators and others

```
corMatrix <- cor(x = differentStudents[,c("Topic_Intention", "Repetitive_Intention", "Concentration", "
corrplot(corMatrix, method = "circle")
```



Analyzing the behavior of specific studen

Define a new coordinate system required for radar chart

```
coord_radar <- function (theta = "x", start = 0, direction = 1)
{
  theta <- match.arg(theta, c("x", "y"))
  r <- if (theta == "x")
    "y"
  else "x"
  ggproto("CordRadar", CoordPolar, theta = theta, r = r, start = start,
          direction = sign(direction),
          is_linear = function(coord) TRUE)
}
```

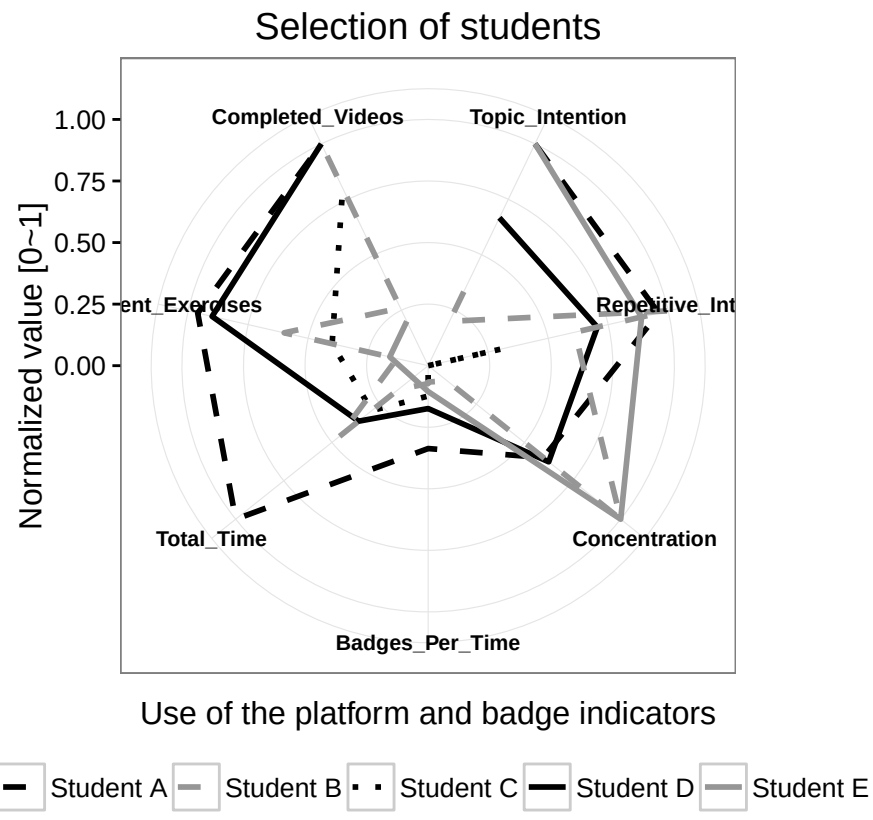
Change the format of data for the visualization

```
selectedData <- allStudents[,c("Topic_Intention", "Repetitive_Intention", "Concentration", "Badges_Per_T

# rescale all variables to lie between 0 and 1
selectedDataScaled <- as.data.frame(lapply(selectedData, ggplot2:::rescale01))

# Adding ID of students as variable
selectedDataScaled$User_id <- allStudents$User_id
```

```
# Melting to long format data
as.data.frame(melt(selectedDataScaled,id.vars="User_id")) -> meltSelectedDataScaled

# interestingStudents variable contains the subset of students that we want to plot
meltSelectedDataScaled <- subset(meltSelectedDataScaled, User_id %in% interestingStudents)
```

Once the data is in the correct format we can build the radar chart to compare the different students, we use ggplot2

```
selected_plot <- ggplot(meltSelectedDataScaled, aes(x = variable, y = value, color= User_id, linetype =
    geom_path(aes(group = User_id), size = 1) + coord_radar() +
    theme_bw() + theme(strip.text.x = element_text(size = rel(0.8), face = "bold"),
        axis.text.x = element_text(size = rel(0.8) , face = "bold"),
        legend.position = "bottom") +
    labs(title = "Selection of students", x="Use of the platform and badge indicators", y = "Normalized
    scale_linetype_manual(values=c(1,3,2,1,2), name="",
                          breaks=interestingStudents,
                          labels=c("Student A", "Student B", "Student C", "Student D", "Student E")) +
    scale_color_manual(values=c('black','black',
                                'black', '#969696',
                                '#969696'), name="",
                       breaks=interestingStudents,
                       labels=c("Student A", "Student B", "Student C", "Student D", "Student E"))

print(selected_plot)
```



Clustering students by their badge indicators

We peform a Two Step Cluster analysis using SPSS software, and store the cluster clasification within 'Cluster_TwoSteps' variable. Now we present some descriptive summary of the metrics for each cluster.

```
# Metrics for cluster analysis
metricsCluster <-  c("Topic_Intention", "Repetitive_Intention", "Concentration", "Badges_Per_Time", "To

# Summary for Cluster 1
summary(subset(differentStudents, Cluster_TwoSteps == 1)[,metricsCluster])
```

```
##  Topic_Intention  Repetitive_Intention Concentration    Badges_Per_Time
##  Min.   : 20.00   Min.   :  0.00       Min.   : 18.64   Min.   : 1.290
##  1st Qu.: 33.33   1st Qu.: 52.20       1st Qu.: 47.09   1st Qu.: 3.015
##  Median : 50.00   Median : 61.98       Median : 59.45   Median : 4.445
##  Mean   : 57.48   Mean   : 62.44       Mean   : 63.89   Mean   : 5.201
##  3rd Qu.: 67.50   3rd Qu.: 75.00       3rd Qu.: 83.67   3rd Qu.: 6.590
##  Max.   :100.00   Max.   :100.00       Max.   :100.00   Max.   :19.790
##    Total_Time     Proficient_Exercises Completed_Videos
##  Min.   :  43.0   Min.   :  6.00       Min.   :  0.00
##  1st Qu.: 204.0   1st Qu.: 20.00       1st Qu.: 15.50
##  Median : 402.5   Median : 43.00       Median : 44.00
##  Mean   : 489.2   Mean   : 46.88       Mean   : 46.76
##  3rd Qu.: 650.5   3rd Qu.: 63.75       3rd Qu.: 81.50
##  Max.   :2458.0   Max.   :100.00       Max.   :100.00
```

```
# Summary for Cluster 2
summary(subset(differentStudents, Cluster_TwoSteps == 2)[,metricsCluster])
```

```
##  Topic_Intention   Repetitive_Intention Concentration     Badges_Per_Time
##  Min.   : 0.0000   Min.   : 0.0000      Min.   : 0.0000   Min.   :0.000
##  1st Qu.: 0.0000   1st Qu.: 0.0000      1st Qu.: 0.0000   1st Qu.:0.000
##  Median : 0.0000   Median : 0.0000      Median : 0.0000   Median :1.110
##  Mean   : 0.3728   Mean   : 0.4698      Mean   : 0.5255   Mean   :1.183
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000      3rd Qu.: 0.0000   3rd Qu.:1.770
##  Max.   :33.3300   Max.   :25.0000      Max.   :54.6600   Max.   :6.380
##    Total_Time    Proficient_Exercises Completed_Videos
##  Min.   :  0.0   Min.   : 0.000       Min.   : 0.00
##  1st Qu.: 25.0   1st Qu.: 0.000       1st Qu.: 2.00
##  Median : 66.0   Median : 0.000       Median : 9.00
##  Mean   :125.8   Mean   : 1.718       Mean   :18.93
##  3rd Qu.:182.0   3rd Qu.: 3.000       3rd Qu.:22.00
##  Max.   :839.0   Max.   :36.000       Max.   :99.00
```

```
# Summary for Cluster 3
summary(subset(differentStudents, Cluster_TwoSteps == 3)[,metricsCluster])
```

```
##  Topic_Intention  Repetitive_Intention Concentration    Badges_Per_Time
##  Min.   : 0.000   Min.   : 30.43       Min.   : 0.000   Min.   :0.440
##  1st Qu.: 0.000   1st Qu.: 50.00       1st Qu.: 0.000   1st Qu.:1.660
##  Median : 0.000   Median : 66.67       Median : 0.000   Median :2.570
##  Mean   : 3.204   Mean   : 70.90       Mean   : 1.936   Mean   :2.789
```

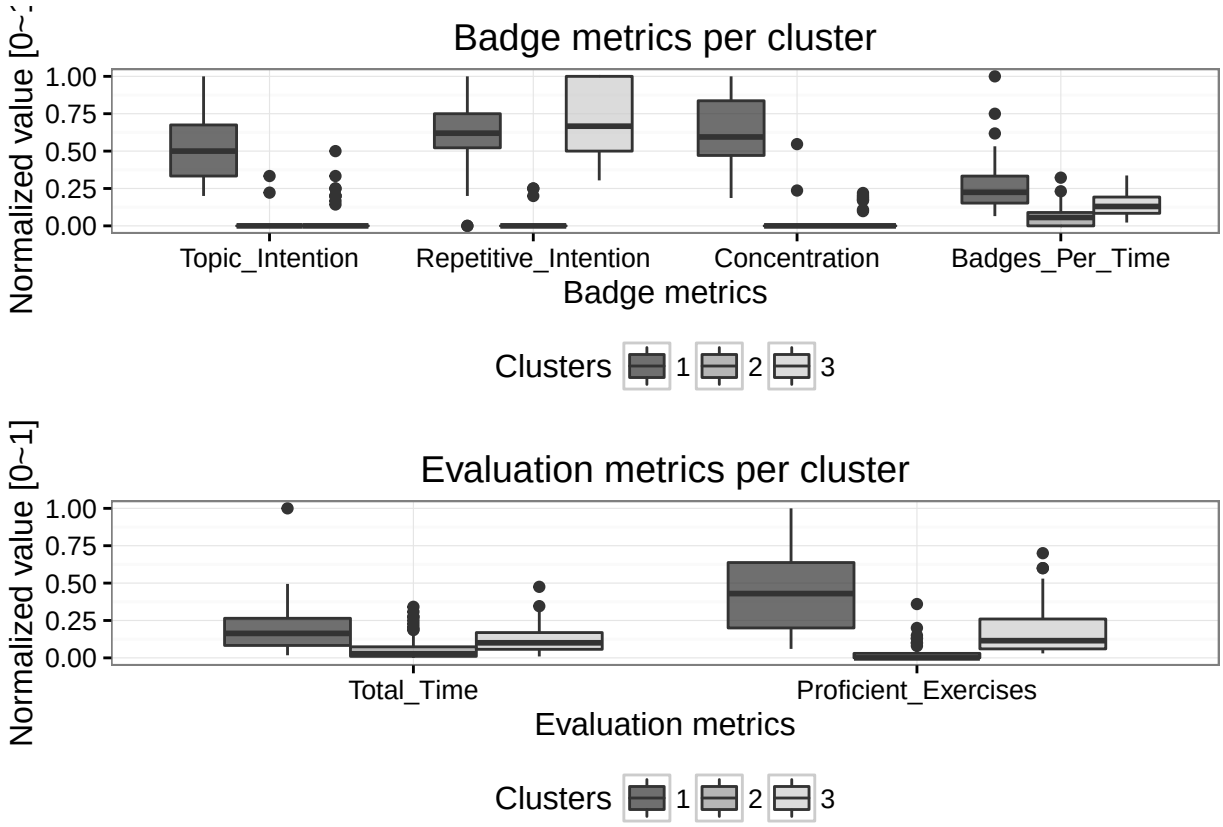
```
##  3rd Qu.: 0.000   3rd Qu.:100.00       3rd Qu.: 0.000   3rd Qu.:3.815
##  Max.   :50.000   Max.   :100.00       Max.   :21.970   Max.   :6.670
##    Total_Time     Proficient_Exercises Completed_Videos
##  Min.   :  23.0   Min.   : 3.00        Min.   : 0.00
##  1st Qu.: 140.5   1st Qu.: 6.00        1st Qu.:12.00
##  Median : 247.5   Median :11.50        Median :30.00
##  Mean   : 314.8   Mean   :17.91        Mean   :39.07
##  3rd Qu.: 415.8   3rd Qu.:26.00        3rd Qu.:68.00
##  Max.   :1168.0   Max.   :70.00        Max.   :99.00
```

Prepare and transform the format of data for the visualizations

```
# Select variables we want to use for plot
differentSelectedData <- differentStudents[,c("User_id", "Course", "Cluster_TwoSteps", "Topic_Intention

# Rescale all variables to lie between 0 and 1
maxs <- apply(differentSelectedData[, !names(differentSelectedData) %in% c("User_id", "Course", "Cluste
mins <- apply(differentSelectedData[, !names(differentSelectedData) %in% c("User_id", "Course", "Cluste

differentSelectedData[, !names(differentSelectedData) %in% c("User_id", "Course", "Cluster_TwoSteps")] <

badge.metrics <- c("Topic_Intention", "Repetitive_Intention", "Concentration", "Badges_Per_Time")
evaluation.metrics <- c("Total_Time", "Proficient_Exercises", "Video_Progress")

meltDifferentStudents <- melt(data = differentSelectedData, id.vars = c("User_id", "Course", "Cluster_Tw
```

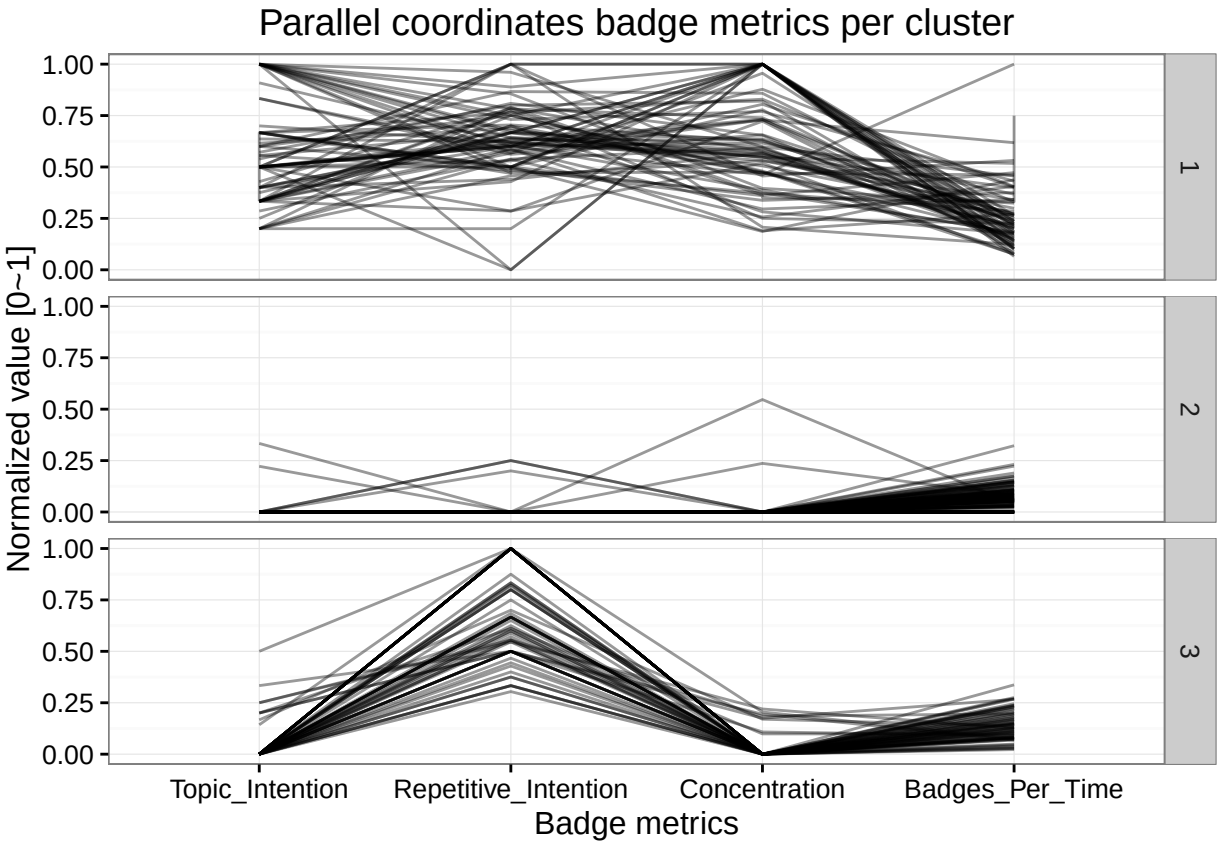
Create the two boxplots facetted by cluster. One for badge metrics and the other one with evaluation metrics

```
p_badge_metrics <- ggplot(data = subset(meltDifferentStudents, Indicator %in% badge.metrics))  +
    geom_boxplot(aes(x=Indicator, y = value, fill = factor(Cluster_TwoSteps)), width = 1, alpha= .7) + t
    labs(x = "Badge metrics", y = "Normalized value [0~1]", title = "Badge metrics per cluster") + scale

p_evaluation_fields <- ggplot(data = subset(meltDifferentStudents, Indicator %in% evaluation.metrics))
    geom_boxplot(aes(x=Indicator, y = value, fill = factor(Cluster_TwoSteps)), width = 1, alpha= .7) + t
    labs(x = "Evaluation metrics", y = "Normalized value [0~1]", title = "Evaluation metrics per cluste

# create a grid with the two boxplots
grid.newpage()
pushViewport(viewport(layout = grid.layout(2, 1, heights = unit(c(5, 5), "null"))))
print(p_badge_metrics, vp = viewport(layout.pos.row = 1, layout.pos.col = 1))
print(p_evaluation_fields, vp = viewport(layout.pos.row = 2, layout.pos.col = 1))
```



Finally, make a parallel coordinates visualization for the badge metrics facetted by cluster

```
p_parallel <- ggplot(aes(y = value, x = Indicator, group = User_id), data = subset(meltDifferentStudents
    facet_grid(Cluster_TwoSteps ~ .) + theme_bw() + labs(x = "Badge metrics", y = "Normalized value [0~
print(p_parallel)
```


Parallel coordinates badge metrics per cluster
Normalized value [0–1]
1.00
0.75
0.50
0.25
0.00
1
2
3
Topic_Intention
Repetitive_Intention
Concentration
Badges_Per_Time
Badge metrics