animbook: An R Package for Visualize …

by Krisanat Anukarnsakulchularp

Abstract An abstract of less than 150 words.

1 Introduction

With the increasing computational powers, animation with large amounts of data is now become more accessible than ever. With the inspiration from a New York Times article, "Extensive Data Shows Punishing Reach of Racism for Black Boys". The animation portrays the trajectory of how boys with different demographics could land up in society. With this plot in mind, the paper proposed a new communication tool to help capture the movement of the observation for each different group over the specific period.

This paper will focus on designing and validating the new visualization using a nested model (Munzner (2009)) to offer an animation tool for conveying the observation trend in the given period. The cognitive theory of multimedia learning (Mayer (2005)) will be considered during the designing period. This theory employs the structure and the strategies that will help the learner learn more effectively.

The challenge of the New York Times article animation was the amount of code needed to reproduce. Additionally, the business data may or may not have the same information as the one in that article. It establishes a need to create tools for generalizing the animation to a wider range of data while still maintaining the main objective.

The structure of this paper will consist of four main section. The first section is domain problem and data characterization, identifying the problems the visualization tools try to solve. Secondly, the operation and data type abstraction, the description of operations, and the types of data required for the next stage. The third section is visualization and interaction design, designing the visuals and refining the tools. In the forth section, package design, develop the structure of the packages and how to implement the final visualize design into the R package. Lastly, the package usage. The checking of the threat and validating of that threat will be carried out throughout the entire process of designing the animbook package.

2 Problem and data characterization

To look for the flow between group, one of the most common tools we look at is sankey chart. It represent the movement of the values from one set to another. However, with the static sankey, the only information presented are what is the proportion of the movement, and which states the quantities move to, this often enough in explaining the flow and helps making the decision necessary. To present more information, color could be used to represent the demographic of the group. What if the information that are interesting lies in the times domain, how could the plot captured that information?

!!general sankey plot and sankey with color

The technique that is created by Hans, Ola, and Anna Rosling in 2003 is called Gapminder Trendalyzer. Its add an additional dimensions which are the bubble size, animated over changes in a fourth dimension, time. The animation in this tools has created a new way for people to analyse and visualize the trend.

!!read effectiveness of animation in trend visualization, then add more information to above paragraph

The osiris data from Bureau van Dijk, contains an information on listed, and major unlisted/delisted companies across the world. This is one of an example of accounting data, where the objective is to observed the movement of the company between the quantile and how japan companies are ranked compared to others.

3 Operation and data type abstraction

When working with the raw data, the data pre-processing is one of the important steps we need to take before further analysis is taken. The function in this package will offer some flexibility in the types of the variable it accepted. However, there are still some restrictions the user needs to follow. In this section, it will provide what are the accepted format and examples for the user the followed to reproduce the final animate plot.

The data set that will be used in this section and in all of the examples of this journal is collected from Bureau van Dijk. The data set contained 30,000 rows and 94 variable of information on listed, and major unlisted/delisted companies across the world. From the raw data, we are only interested in the ranking of the companies. The cleaned version of this data set is included in the package which only contained the relevant variables to reproduce the animated plot.

4 Visualization and interaction design

5 Package design

In designing the package for reproduces the NYtimes animated plot, the package end up with a three steps process in recreates the animation. The first step which is prepared the data into the right format for the plot function. The next stages is to create a ggplot object which then can be inputted into animation function. The last step is to added the animation settings to the ggplot object so that the gganimate::animate() function. The reason for this three steps process is that it allows the user who does not have a lot of experience to reproduces the animation.

As per motivations, the main focus for this package was to look at movement between the percentile group. Thus the anim_prep() function first purposes was to be able to take any numerical values and output the percentile group of the observations. The algorithms behind the group splitting are quantile and cut function. The quantile function from the stats R packages takes in the numeric vector and output the corresponding quantiles to the given probabilities. The output from the quantile function can then be used as the breaks argument for the cut function that is from the base R packages. The concerns of this method is that the users might not always interests in ranking the observations.

The next logical design is to allows for different scaling. As mentioned in the paragraph above the first scaling is “rank”. The function then expanded the function to allows for the “absolute” scaling. This scaling is to assign the quantile groups based on the absolute values scales. The default for the function is to breaks the group equally using seq function. The input to the seq function are taken by the min and max of the values and increment by the equal steps depend on number of group of interests.

The algorithms mentions so far only breaks the quantile equally, but that is not always the cases as the users may interested in the group that are not equally breaks.

6 Package Usage

Data preprocesing step

This first step can be done using the anim_prep() function. This function required that the users data contained id which is used to uniquely identified each individual observations, values which is numerical values to be used to group the observation together, and time, the variable that distinguish the values between the time and id. The users will need to understands what data structure can be input into the function. This package included the osiris data as an examples to the users.

7 `{r} # animbook <- anim_prep(osiris, ID, sales, year) #`

There are an additional options that allow for more customizations to the data structure or the plot.

- label: group labeling.
- ngroup: number of groups we want to split the values into.
- breaks: the group bins size (prototype)
- group_scaling: the grouping variable for the bins calculations.
- color: the variable used to color the observations.
- time_dependent: logical. Whether we want the observations to start at the same time or not.
- scaling: the scaling method, either `rank` or `absolute`.
- runif_min: minimum value for random addition to frame numbers
- runif_max: maximum value for random addition to frame numbers

The function can calculate four different scales using these options.

```
8 {r} # # rank scaling # rank_scaling <- anim_prep(data = osiris,
  id = ID, values = sales, time = year) #  # # absolute scaling #
  absolute_scaling <- anim_prep(data = osiris, id = ID, values =
  sales, time = year, #                                scaling =
  "absolute") #  # # rank scaling by group # rank_group_scaling
  <- anim_prep(data = osiris, id = ID, values = sales, time =
  year,  #                                    group = country)
  #  # # absolute scaling by group # absolute_group_scaling <-
  anim_prep(data = osiris, id = ID, values = sales, time = year,
  #                               group = country, scaling =
  "absolute") #  # rank_scaling #
```

This function will return `animbook` object containing a list of the formatted data and settings.

Plotting the data

After preparing the data, we can not plot it. There are three plots available in this package:

- `kangaroo`, which plots the observation's movement over time.
- `wallaby`, which subset the plot to either `top` or `bottom` and see which group they are in after the observational period.
- `funnel_web_spider`, which is a faceted plot by time variable.

```
9 {r} # label <- c("Top 20%", "20-40", "40-60", "60-80", "80-100",
  "not listed") #  # animbook <- anim_prep(data = osiris, id = ID,
  values = sales, time = year, color = japan, label = label) #  #
  # kangaroo plot # kangaroo_plot(animbook) #  # # wallaby plot
  # wallaby_plot(animbook) #              # # funnel web spider
  plot # funnel_web_plot(animbook) #
```

The `kangaroo` and `wallaby` plots can be animated using the function of the next stage. `funnel_web_spider` only supported static plot. We can also choose whether we want to animate the plot using gganimate or plotly.

Animating the plot

To animate the plot, we need to save the plot into an object, which then can be passed on to the function.

```
10 {r} # animbook <- anim_prep_cat(data = aeles, id = id, values
   = party, time = year, color = gender, time_dependent = FALSE)
   #  # p <- wallaby_plot(animbook) #  # p2 <- anim_animate(p) #
   # gganimate::animate(p2) #
```

Introduction

Some ideas:

- The animated plot has not been used often in the business data presentation
- We want to see the movement of our variable of interest
- Business data: Accounting data (sales), Marketing data (customer interest), data analysis languages data
- Help with storytelling
- Inspiration: The New York Times (Extensive data shows punishing reach of racism for black boys). The animation plot shows what demographic group between white and black boys where would end up in society. It could be adapted to show, based on the demographic of the company,

sees the movement of where they would end up. In the marketing data, the demographic of the customers could tell us the story of what product they preferred.

Data and Process

Some ideas:

- The data will need a following structure for it to be ready to plot: time index(time variable for the gganimate), x-axis for ggplot(this does not need to be the same as time index), key(unique identifier), rank(y-axis, factor variable), group(optional).
- The user can input any numerical, categorical and factor variable into the prep_anim function which will return the data in the format that the anim_plot function accepted. The x-axis and time index does not need to be the same as this allows for the user to reproduce the nytimes plot which is not time dependent. For the rank variable, no matter what the type of variable is in, it will return numerical variable.

When working with the raw data, the data pre-processing is one of the important steps we need to take before further analysis is taken. The function in this package will offer some flexibility in the types of the variable it accepted. However, there are still some restrictions the user needs to follow. In this section, it will provide what are the accepted format and examples for the user the followed to reproduce the final animate plot.

The data set that will be used in this section and in all of the examples of this journal is collected from Bureau van Dijk. The data set contained 30,000 rows and 94 variable of information on listed, and major unlisted/delisted companies across the world. From the raw data, we are only interested in the ranking of the companies. The cleaned version of this data set is included in the package which only contained the relevant variables to reproduce the animated plot.

- The data structure
- The steps for the user
- How does the function processes the data

References

Mayer, Richard E. 2005. "Cognitive Theory of Multimedia Learning." In *The Cambridge Handbook of Multimedia Learning*, edited by RichardEditor Mayer, 31–48. Cambridge Handbooks in Psychology. Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.004.

Munzner, Tamara. 2009. "A Nested Model for Visualization Design and Validation." *IEEE Transactions on Visualization and Computer Graphics* 15 (6): 921–28. https://doi.org/10.1109/TVCG.2009.111.

Krisanat Anukarnsakulchularp
Monash University
Faculty of Business and Economics
Melbourne, Australia
ORCiD: 0009-0008-5638-7124
kanu0003@student.monash.edu