清華大學

Tsinghua University

Department of Computer Science and Technology

Advanced Computing

Assignment 1

Name: Sahand Sabour

Student ID: 2020280401

Module: Advanced Network Management

Introduction

With the significant abundance of data and database systems in the recent years, data visualization, which refers to the graphical representation of data, has become an essential tool of data analysis and interpretation as it provides an accessible way to view and acknowledge data patterns, trends, etc. By many, good data visualization is considered as a beneficial skill that takes time and dedication to master. Accordingly, data pre-processing also plays an essential role in modern data analysis as it provides a much better and clearer foundation for the analysis. In this assignment, a given data-set is analyzed and pre-processed, and its attributes are visualized in different graphs to investigate the trends and patterns that at first glance, may not be easily discovered. Hence, the aim of this assignment is to introduce students to different data pre-processing and visualization techniques and approaches while cultivating their skill of beneficial data analysis. In this report, a brief background of the assignment and its tasks are mentioned, and the corresponding results, discussions, and answers are provided.

Background

The provided data-set contained 2 weeks of search logs from a large search engine, where each day of logs was saved in a separate CSV file. Each row in the log file has the following attributes:

```
Timestamp, #Images, UA, Ad, ISP, Province, PageType, Tnet, Tserver, Tbrowser, Tother, SRT
```

Accordingly, the meaning of each of the above attributes was provided as below:

- **Timestamp**: the Unix query timestamp.
- **#Images**: the number of images embedded in the response.
- **UA (User Agent)**: the type of the user's browser where the query submitted from.
- **Ad**: whether response contains ads or not, "AD" for yes and "noAD" for not.
- **ISP**: the Internet Service Provider.
- **PageType**: whether the page is loaded synchronously or asynchronously.
- **Tnet**: the page transmission time over the network.
- **Tserver**: the server-side processing time of the query.

- **Tbrowser**: the DOM parsing time of the browser.
- **Tother**: the remaining time for obtaining other embedded elements.
- **SRT**: search response time (ms), which is the sum of the above four Ts.

Tasks and Results

The given assignment tasks and corresponding responses are provided respectively below.

1) Calculate the average SRT of every 10 minutes, and plot the SRT with a line chart (x axis for date time and y axis for the average SRT).

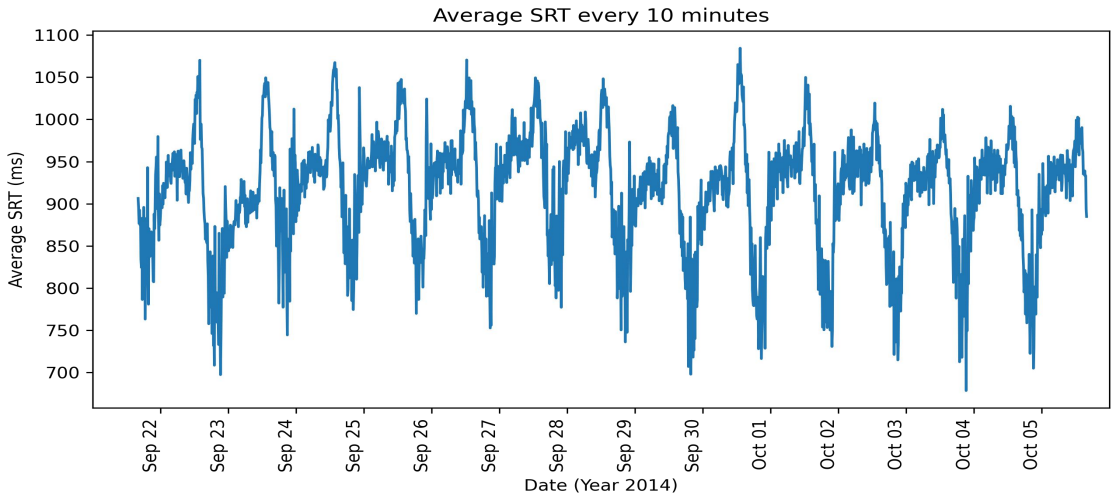

Figure 1: Average SRT per 10 Minutes line chart

The required line chart is delineated in the above figure (Figure 1). Line charts are mainly used to convey a connection between a series of data points with their time of occurrence as a continuous line. That is, line charts best describe a variable's pattern of change in the given time range. By creating the line chart for the average SRT per 10 minutes, it can be observed that the maximum average SRT would occur during approximately the same time of each day: evening to the end of the day. Moreover, it can also be observed that the lowest average SRT occurs approximately after midnight for each of the investigated 14 days. Hence, based on these observations, it could be derived that the SRT for this search engine experiences its lowest value at some time after midnight and highest value before the day's end while this value increases throughout the day.

2) Calculate the average of each SRT component of every 10 minute, and plot the four SRT components together with a stacked area chart (x axis for date time and y axis for time) and also a 100% stacked area chart (y axis for the percentage).

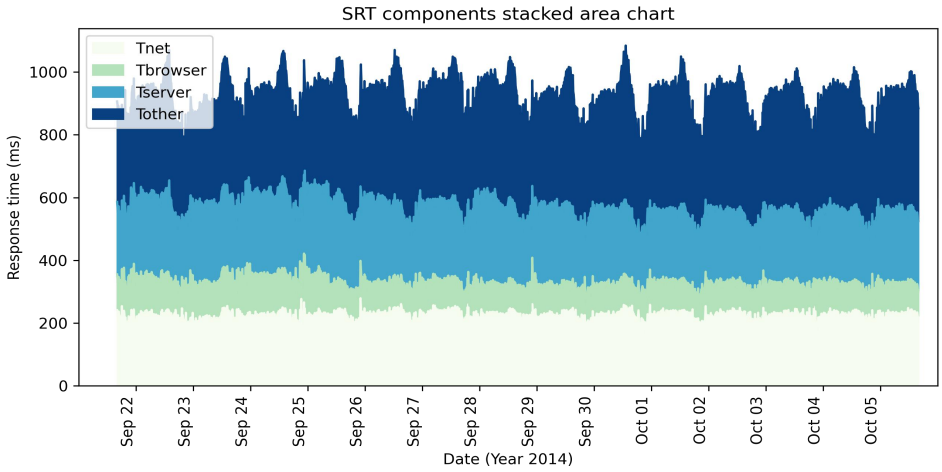

Figure 2: Average SRT components per 10 Minutes stacked area chart

Stacked area charts are beneficial when an attribute consisting of different components is to be investigated as they demonstrate the evolution and variation of the multiple different components on the same graph. These types of charts also highlight the importance and significance of each component in respect to other components. For instance, in the given data-set, SRT consists of four components: Tnet, Tbrowser, Tserver, and Tother, and Figure 2 displays the stacked area chart for these components. As displayed in this figure, it can be derived that Tother is the most significant component of SRT with Tserver being the second significant component. That is, the value of SRT is mostly influenced by Tother.

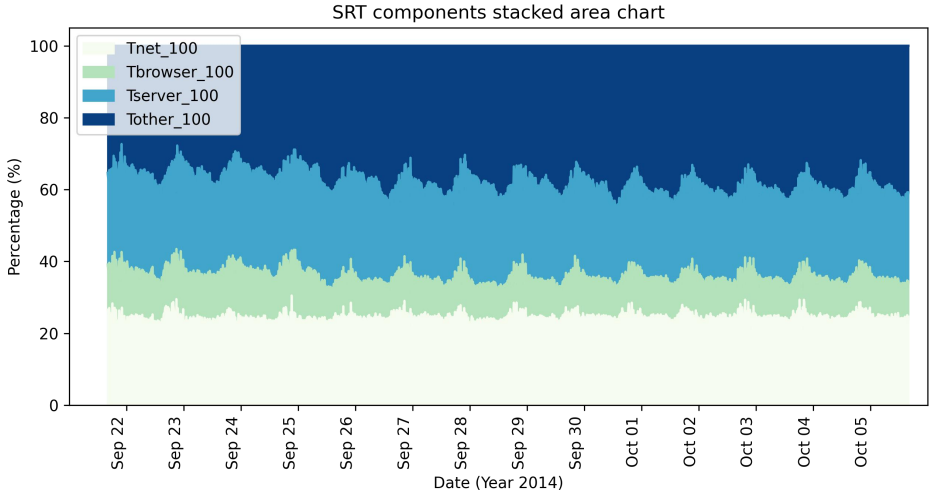

Figure 3: Average SRT components per 10 Minutes 100% stacked area chart

In order to further highlight the significance of each component in the overall perspective, 100% stacked area charts could be used (Figure 3). 100% stacked area chart is the normalized version of the stacked area chart, where the sum of each group at each position is 100%. Hence, based on Figure 3, it can be more clearly observed that Tother is the component that contributes the most to the value of SRT (approximately 30%).

It should be noted that, unlike the line chart, the stacked area chart may not be a good representative of the variations over time as the displayed values of the components in the middle is shown as the sum of the component's own value and all its below components.

3) Plot the CDF (Cumulative distribution function) chart of SRT.

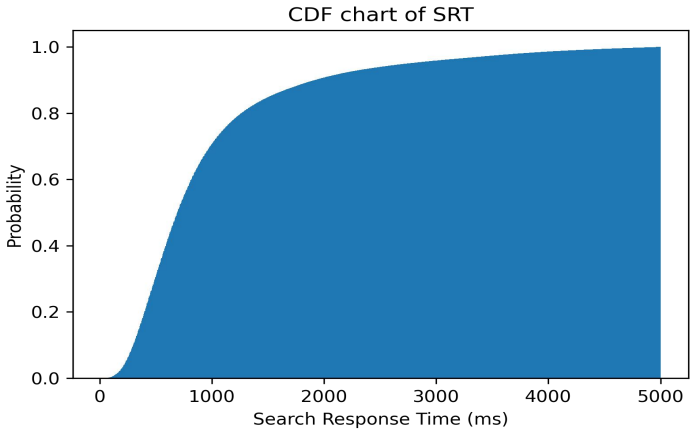

Figure 4: SRT CDF chart

Cumulative Distribution Function (CDF) charts demonstrate the probability that a given variable X has value x. Hence, these charts are mainly used to determine the probability that a random observation is less than or equal to a specific value; more specifically, how the data is distributed. For instance, Figure 4 illustrates the CDF chart of SRT. As shown in the figure, it can be observed the probability at the point SRT = 5000 ms is 1; that is, all the recorded SRT values are less than or equal to 5000 ms. Moreover, the probability of SRT $\leq 1000 = 0.7$, which means that there's a 70% probability that a query experience an SRT of less than or equal to 1000 ms. CDF charts can also be used to calculate the probability that a given value is between a given range. For example, based on Figure 4, there's a 20% probability that a query's SRT is between 1000 ms and 2000 ms. Lastly, it can also be derived that although the maximum value for SRT is 5000 ms, it is barely probable that any query experience this value as queries have a 90% chance of having SRT $\leq$ 2000 ms

4) Plot the CDF chart of #Images.

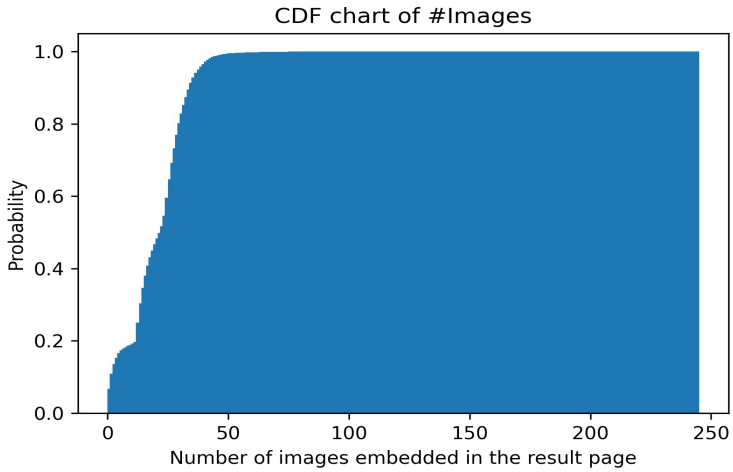

Figure 5: #Images CDF chart

Similar to the previous task, Figure 5 demonstrate the CDF chart for the number of images that are embedded in the fetched webpage. According to this chart, almost all the queries include less than 50 images in their responses. Moreover, there's a 50% chance that a result page includes 25 or less images.

5) Count the number of queries (also called page views or PVs) of each minute, and plot the minute-level PVs with a line chart (x axis for date time and y axis for the PVs).

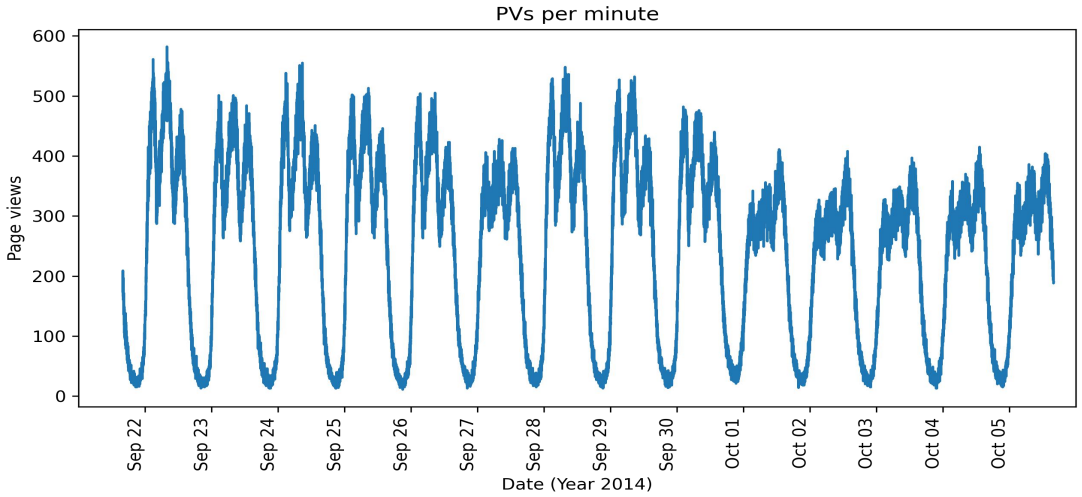

Figure 6: Page views per minute line chart

Similar to the first task, the line chart for Page Views (PVs) per minute is illustrated in the above figure (Figure 6). In addition, it can be observed that the number of queries are considerably larger in the period between noon and evening in contrast to other times in the

day. It can also be observed that the number of queries are significantly low starting from midnight till the morning. By mixing the conclusions from the first task, it can be derived that there exists a correspondence between the average SRT and number of queries; where the periods with higher number of queries have larger average SRT. The point to have in mind is that the processing increments in the first task are 10 minute intervals whereas we use 1 minute intervals in this task. Hence, the derived conclusion is intuitive and is not established by completely correlating the two plots.

6) Count the PVs of each province, and plot it with a histogram chart (x axis for province and y axis for PVs).

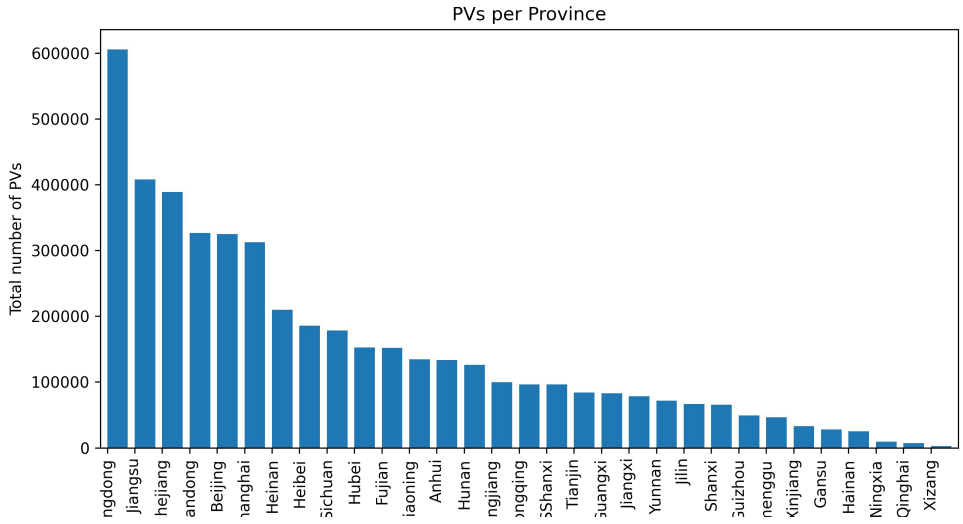

Figure 7: Page views per province bar chart

Histograms are commonly used to show the distribution of numerical data. That is, histograms plot ranges of data grouped together. However, when discussing the number of queries from different provinces, we are required to divide the data into different categories (provinces in this case). Therefore, we should plot the bar chart instead of the histogram. The bar chart regarding Page Views (PVs) per province is illustrated in the above figure (Figure 7). Upon observing this chart, it can be noticed that the queries coming from a number of provinces are considerably larger in comparison with other provinces. For instance, the largest number of queries are sent from Guangdong province. Hence, it can be derived that this province plays an important and rather decisive role in the analysis of this data-set as a significant part of the entries have come from this province.

7) Count the PVs of each UA, and plot it with a pie chart (show the percentages).

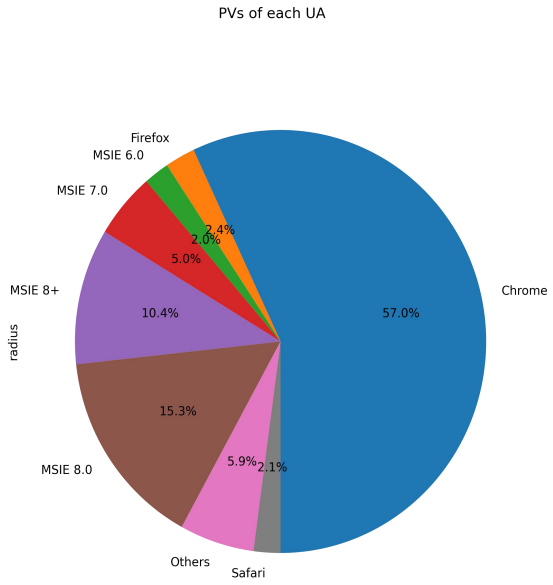

Figure 8: Page Views for each User Agent Pie Chart

Pie charts show the relationship between different percentages. The idea is that the sum of this percentages would be 100% and that is the whole area of the circle. Hence, pie charts are mainly used to highlight the differences in proportions of an attribute. For instance, as shown in Figure 8, the pie chart for the PVs based on different User Agents (UAs) or browsers demonstrates that the majority of the queries (57%) were made from the Chrome browser. In addition, it also shows that the least amount of queries came from MSIE6.0, which is to be expected as this is an outdated browser. Developers could use this chart to identify the browser with the most queries and focus on the services provided on this browser comparatively more than the others.

8) What are the differences among those charts (How to decide which one to use)

In this assignment, line, stacked area, 100% stacked area, histogram, bar and pie charts were discussed. Accordingly, brief introduction of each type of chart alongside its use-case was provided.

To summarize what was previously mentioned, line charts display different values of a variable corresponding to their time of occurrences. Hence, if the history of variations in an attribute is being investigated, line charts would be used. Line charts are especially useful in

machine learning models that learn the history of a variable to predict future its values.

Stacked area charts and 100% stacked area charts are implemented to highlight the change of several components on the same graph. However, unlike the line chart, only the change of the lowest component in the graph corresponds to its actual value while the other layers demonstrate the sum of all lower layers' values. Therefore, these types of charts would be usually used when the focus is determining the most significant component of an attribute.

Histograms illustrate the distribution of data in intervals and their frequency of occurrences. Hence, they are mainly used for visualizing the distribution of quantitative data in discrete intervals. On the other hand, bar charts are used to display categorical data. Hence, when comparing the appearance of a histogram and a bar chart, the bars in a bar chart are separated from each other while the is no space between the bars of the histogram as they represent intervals rather than values (i.e. categories). Lastly, pie charts display the distributions of percentages and their relationships. Therefore, pie charts are used to illustrate data distributions of continuous data and percentages that are limited to few categories. In situations where the data is discrete and the categories are rather too diverse, histogram or bar charts would be much better choices than pie chart.

9) Describe your experience or findings in doing those jobs. For example, experience of processing the data, observations from the charts, characteristics of the data, potential explanations, and any interesting things you would like to mention.

Personally, I had little experience with processing data and data visualization using Python prior to doing this assignment; although I have wanted to learn it for a long time as I consider highly necessary with today's standards. Therefore, in order to complete the tasks of this assignment, I had to visit many websites and forums for tips and tricks, whether it was about drawing a specific graph, setting labels, changing colors, etc. The observations and potential explanations of the above charts is provided in each task respectively. I believe that there were many more patterns that could have been realized by combining (grouping) and visualizing different attributes together. In addition, a number of attributes such as Ad or PageType and their respective patterns were not investigated in the given tasks. Hence, as a suggestion, I believe that these additions could be made to further improve this assignment.

Conclusion

In this assignment, a data-set containing the logs of a popular search engine for two weeks were explored, analyzed, and visualized via different charts and techniques. It is believed that this assignment provided the students with an invaluable chance to practice and improve their data visualization and pre-processing skills and overall understanding. In conclusion, this assignment is considered as a significant learning opportunity that has successfully encouraged students to do the required research in order to complete the required tasks.

Reference

[1] FOCUS: Shedding Light on the High Search Response Time in the Wild. INFOCOM 2016.