80240663 Advanced Network Management Assignment 1

Tsinghua University

September 23, 2019

Albert Millan – 2019280366

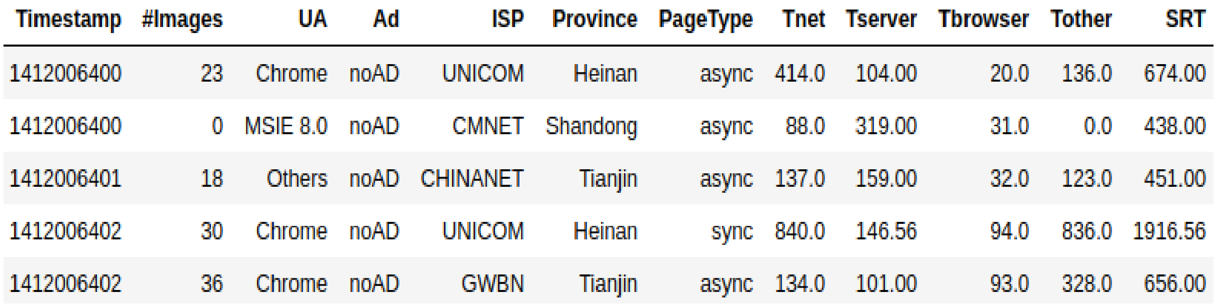
Contents

1 Introduction

This paper aims to provide an analysis of two weeks‘ search engine log data and highlight the key findings. A brief summary of the data is outlined first. The following structure is identical to the one described in the assignment description paper, providing both the required graphs and an explanation at each section. Concluding remarks are presented lastly, summarizing the results and primary insights of the project.

2 Task

Two weeks of query log data was provided, ranging from the 22nd of September to the 6th of October. The attributes include the query submission unix timestamp, number of images embedded in the result page, the user agent, whether the page contained ads, the origin internet service provider of the query, the Chinese province the query was sent from, the page type, the search response time (SRT) and subsequent decomposition as it navigated through the nets components. A small sample of the data is presented in **figure 1**.



Timestamp	#Images	UA	Ad	ISP	Province	PageType	Tnet	Tserver	Tbrowser	Tother	SRT
1412006400	23	Chrome	noAD	UNICOM	Heinan	async	414.0	104.00	20.0	136.0	674.00
1412006400	0	MSIE 8.0	noAD	CMNET	Shandong	async	88.0	319.00	31.0	0.0	438.00
1412006401	18	Others	noAD	CHINANET	Tianjin	async	137.0	159.00	32.0	123.0	451.00
1412006402	30	Chrome	noAD	UNICOM	Heinan	sync	840.0	146.56	94.0	836.0	1916.56
1412006402	36	Chrome	noAD	GWBN	Tianjin	async	134.0	101.00	93.0	328.0	656.00

Table 1: Sample data showing the first five rows.

2.1 Ten Minute Average SRT

The timestamp data shown in table 1 was converted into time series with *datetime64* format and successively aggregated into 10-minute intervals. The mean 10-minute SRT for the given period was then computed and plotted into the line chart shown in figure 1.

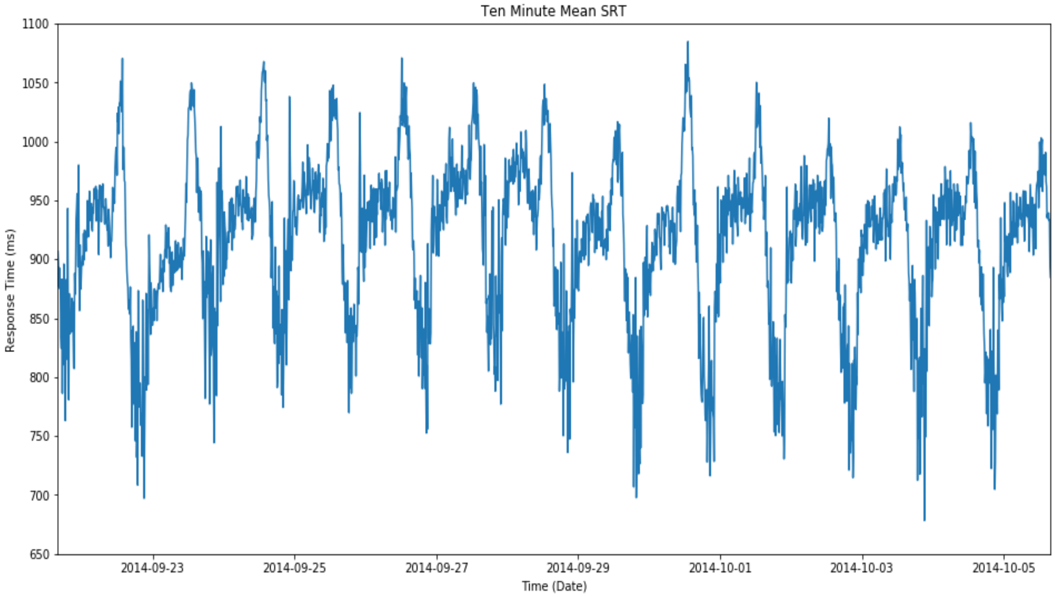


Figure 1: Ten-minute average SRT.

Figure 1 indicates that the SRT data exhibits strong periodicity. SRT is at its lowest during night time, increasing considerably during the morning and reaching its peak on the afternoon and evening. It could potentially be related with the volume of queries received, being at its lowest during night time as people are sleeping, increasing in the morning as people go to work, and having the longest response time in the evening when people have more time to spare online with their devices. A larger volume of queries could potentially slow down the SRT.

It is also worth noticing that the SRT is the slowest on the evening of September 30th, which coincides with the start of the National Holidays in China, and thus a significant volume of queries is to be expected.

2.2 Ten Minute SRT Component Average

Figures 2 and 3 decompose the 10-minute query SRT across its four components. While figure 2 shows the aggregated component response time as a stack of colours adding up to the total SRT, figure 3 shows the percentage of time taken by each component to process the query as a fraction of the total time taken.

A considerable proportion of the total SRT is spent acquiring embedded elements in the page

(TOther) and processing the query in the server (TServer). Also, from figure 3, it can be deduced that the DOM parsing time (TBrowser) is stable (the latter simply maps the processing time fluctuations from the TServer). The volatility of TOther is significantly larger than the other components.This indicates that an increase in the number of images/media would be followed by a larger increase in TOther than that increase expected from the other components.

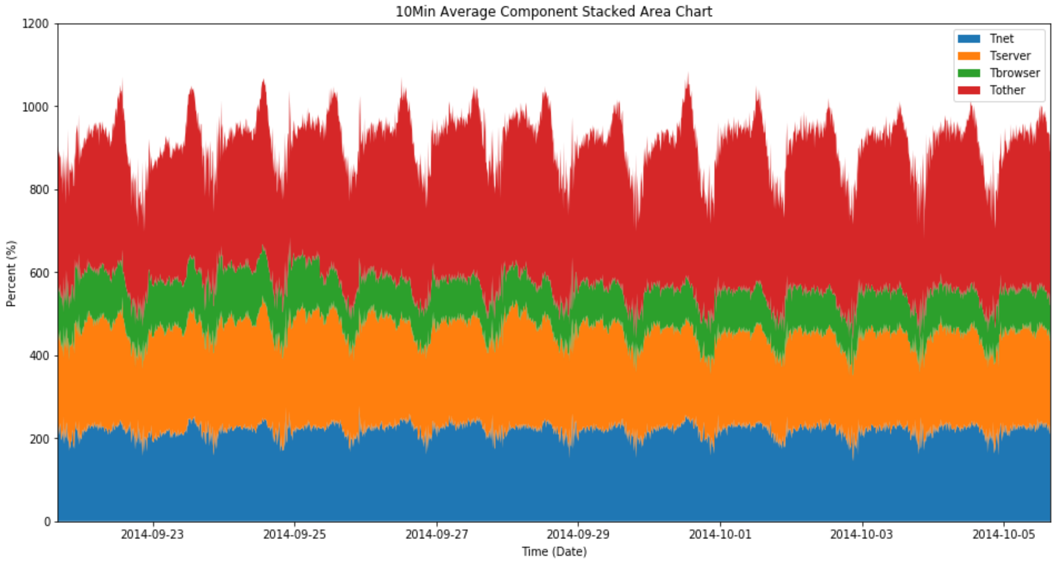


Figure 2: Stacked area chart showing the ten-minute average component SRT.

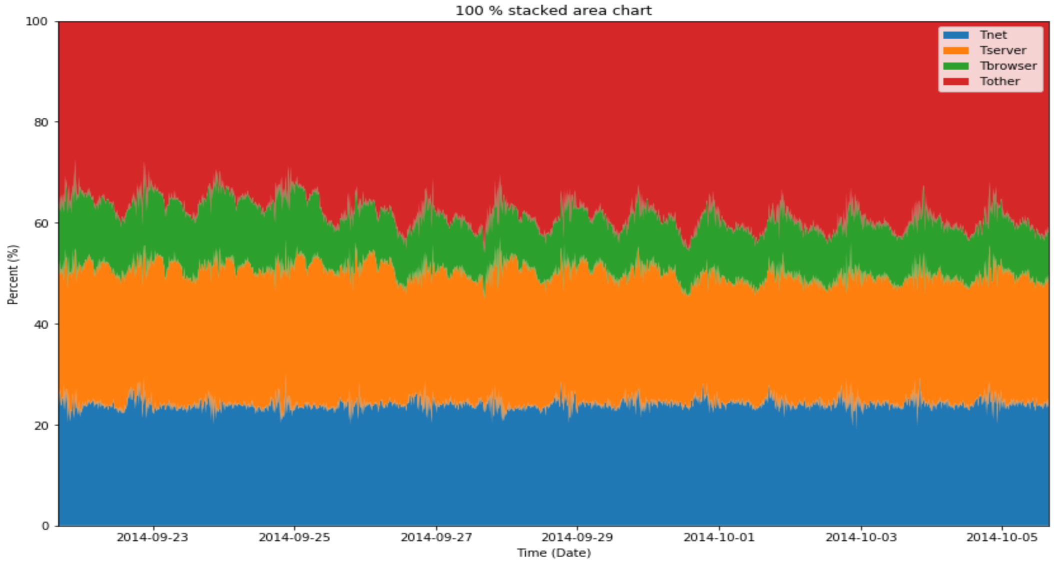


Figure 3: 100% Stacked area chart showing the ten-minute average component SRT.

2.3 SRT Cumulative Distribution Function

The SRT cumulative distribution function (CDF) shows the distribution of SRT across its entries. Figure 4 shows that it takes less than one second to process 73% of the entries, with increasing processing time for the remaining 27%. In some cases, it takes up to five seconds. This can potentially have a negative impact in KPIs such as customer adoption or even profits for companies, as larger waiting time tend to draw users away from the site.

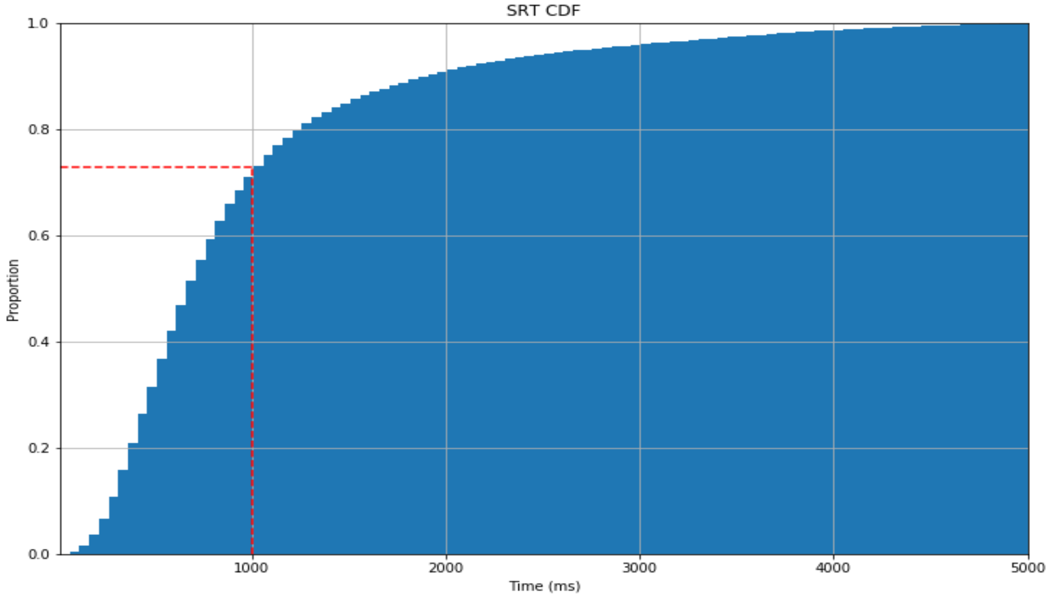


Figure 4: SRT Cumulative Distribution Function.

2.4 Images CDF

The SRT cumulative distribution function (CDF) shows the distribution of the number of images loaded across its entries. Figure 5 indicates that there are up to 12% that loaded no image files, and that the majority of queries required no more than 50 images. However, there were few cases requiring a larger volume of files, being 245 the largest.

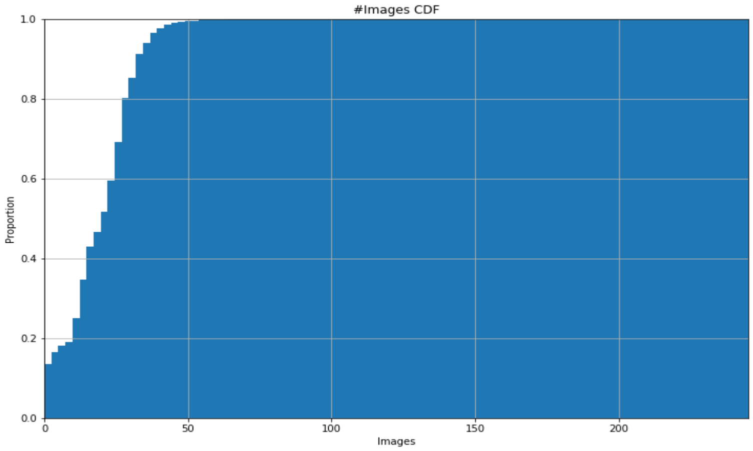


Figure 5: Images Cumulative Distribution Function.

2.5 Minute-Level PVs

The minute-level PVs was computed by aggregating and counting the queries/entries into one-minute intervals. These where then plotted into a line chart. The data from Figure 6 exhibits periodicity. There are three major spikes during daytime, out of which the first and second tend to have larger volumes up until the 30th of September. Thereafter, the trend reverses, being the third spike the one having larger volume. Also, on the latter dates, the spikes aren't steep as in the former, having a lower daily volume of queries overall. At the nighttime, the volume of queries decreases significantly.

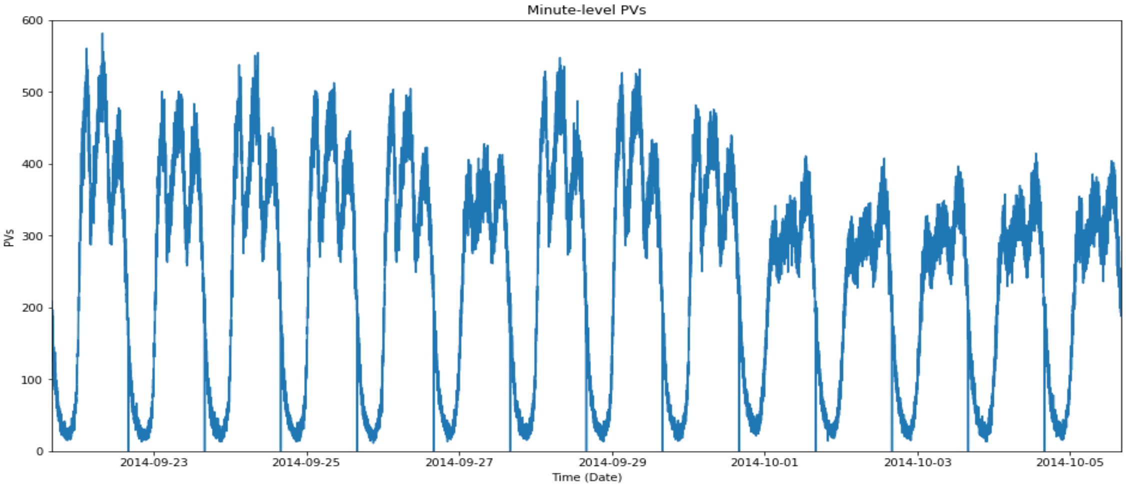


Figure 6: Minute-Level PVs.

2.6 Total PVs Per Province

The volume of queries was segregated and grouped by the province they originated. These were then plotted into a bar chart graph and ordered in descending order to highlight the difference. A significant amount of queries comes from Guandong, followed by Jiangsu and Zhejiang. This information would be extremely useful to derive the location of the server so that it is at the closest point to the locations from which the majority of queries originate.

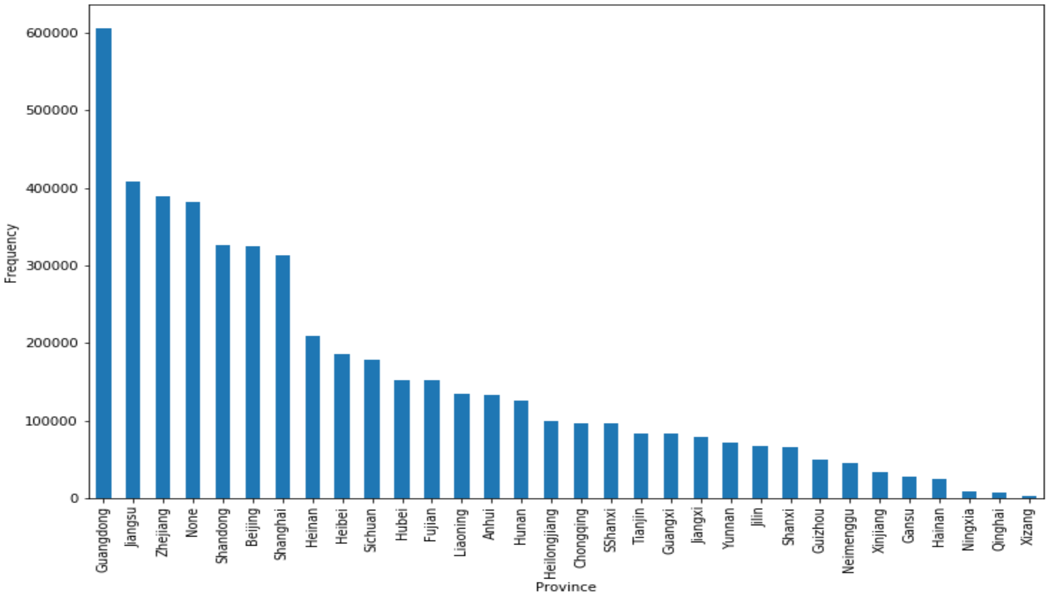


Figure 7: Total PVs across provinces bar chart.

2.7 Percentage of PVs per UA

The pie chart shown in figure **??** outlines the fraction of queries that where sent from each user agent (UA) out of the total number of queries. Up to 57% of the queries originated from Chrome browser, 32% with a MSIE 6.0 or following versions, and only 5% where done in Firefox.

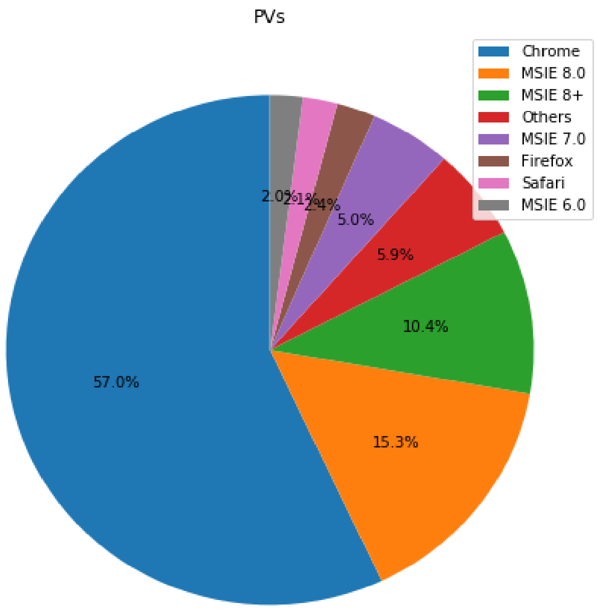


Figure 8: Pie Chart representing the total PVs per UA.

3 Charts

Motives supporting the use of one graph type over the others are presented below:

- **Line Chart**. Convenient to denote how one or many parameters change with respect to time.
- **Stack Chart**. Useful to decompose a variable into several components. Facilitates the comparison between them by stacking one over the others.
- **CDF**. Shows the number/percentage of entries that fall below a certain delimiter of a given attribute. Helpful to determine whether the proportion of acceptable results that fall below the threshold.
- **Bar Chart**. Used to compare the frequency of entries or aggregation of data by attribute across some other categorical data attribute. Provides an insight of the volume of data across each categorical data attributes and by how much do these differ.

- **Pie Chart**. Visually highlights the fraction of a given attribute of each different type, having a fragment for each type. The fragment's size corresponds to the proportion of entries out of the total number of entries.

4 Summary

The hypothesis proposed in section 2.1 explaining the periodicity was incorrect. The relationship between SRT and volume is not strong enough to justify the periodicity denoted in figure 1, therefore, there must be other variables influencing the search response time shifts. Up to 27% of the queries take more than one second to be processed, which is a significantly large amount of time and can draw users away. Another aspect to consider is the impact of media files on the SRT. TOther increases/decreases by a larger factor than the other components on the addition/removal of media files. Regarding PVs, most queries come from the first seven provinces listed in figure 7 starting from the left-hand side, hence the servers should be as close as possible physically possible to these areas in order to minimize network communication costs and enhance response times. Lastly, the majority of users use Chrome to connect with the server. This platform should be prioritized over the others.