Data Science and Engineering (DSE) Record, Volume 1, issue 1.

Analysis and Visualization of Journey Data of Mass Rapid Transit of Bangkok Metropolitan

Sutipong Sutinaraphan[1], Juggapong Nartwichai [2]
[1] Data Science Program, Chiang Mai University, Chiang Mai, Thailand
[2] Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
sutipong_s@cmu.ac.th

Abstract. This independent study aims to create a data transformation process and essential analysis and illustrate visualization from the passenger journey data on mass transit system. In this era, urban and population expansion has caused an incredible concern on urbanization especially infrastructure of the city. Mass public transportation is considered to be the core development for an urban logistic and one of the most important tools to minimize the social disparity. Unfortunately, due to some constrains, one of the most vital data such as passenger travelled data cannot be obtained in ready-to-use form. In this independent study, we will focus on develop the automation process to transform the origin-destination matrix of mass rail transportation into transactional data which will enable the processing of analytical and visualization data further. Furthermore, this study will develop analytical visualization in accordance with the requirements gathered from the executive of the Mass Rapid Transit Authority of Thailand (MRTA). This study has shown significant insights from passengers of the mass rail transit system which will be able to supply as supporting data for yearly and quarterly meeting for MRTA's board of director to improve the ticketing system and most importantly to be a primary supporting data for future public transportation development in Thailand.

Keywords: Trajectory Data, Data Visualization, Mass transit, Journey Data analysis.

1 Introduction

Journey data or so-called trajectory data is one type of time-space data that are very attractive to researcher as it may reveal the pattern of journey or even can help forecast the usage and movement of people in the city [1]. Hence, multidimensional trajectory data of MRT in Bangkok can help both public transportation operators and urban development sector to achieve the goal to be a smart city in other provinces of Thailand. Due to the full expansion of MRT-blue line in year 2019-2020 [2], trajectory data will be continuously feed into the database. In order to study and understand the movement of the people in Bangkok city, trajectory data may play an important role for researcher to discover some useful knowledge which benefit the urban planning and urban developer. Data mining on trajectory data aims for knowledge discovery technique can bring

in to extract knowledge. The analysis and visualization of data would enlighten to the related stakeholder on sustainable development. The study process and methodology of the analysis and visualization are described in Section 2. The analysis and visualization result are shown in Section 3. Next, the discussion and conclusion has been made in Section 4 and Section 5.

2 Study process and methodology

This study adapt methodology mainly from the CRISP-DM framework process which is widely accepted and use for data mining project.[3] The process can be divided into data exploration, data pre-processing, data analysis, data visualization and data presentation to the data owner. The original input data come in form of origin-destination matrix report which cannot be immediately use for analysis and visualization. Previously, the data owner has to do the traditional way to pre-process this mentioned data which this study aims to present with the proposed way (Fig.1). However, the data possessed a completeness and consistency of the data without any error or missing value. Data then processed through pre-processing step to transform the report into transactional data and transform necessary transformed column as well. Then begin the analysis and visualization step by begin with interview the data owner and data user representative to gather all essential requirement. After study details of the requirement and verify the possible analysis scope, then continue with visualization and presentation further.

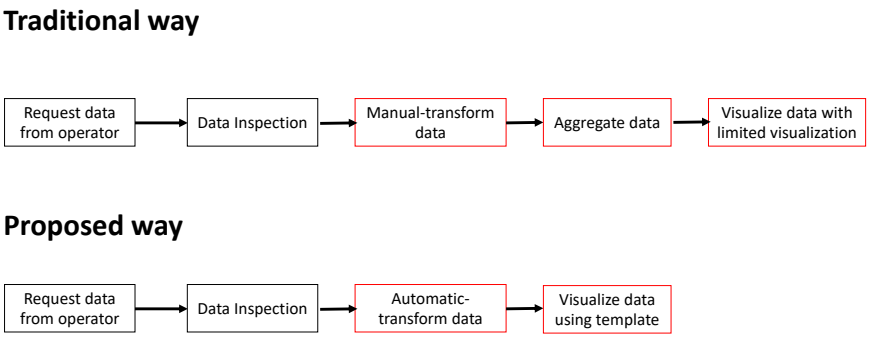

Fig. 1. Traditional data processing VS. Proposed process

3 Analysis and Visualization Result

From the interview and discussion with the data owner team consist of one executive level and two operation level personnel about data analysis requirement, we can conclude their concrete analysis requirement into 3 aspects:

1. Visualization that represent weekday and weekend journey pattern. This is to understand the journey pattern difference between weekday and weekend and to monitor

any adjustment and recommendation that they should suggest to the operator. Fig.2 illustrate the total difference pattern between weekday and week hour journey.

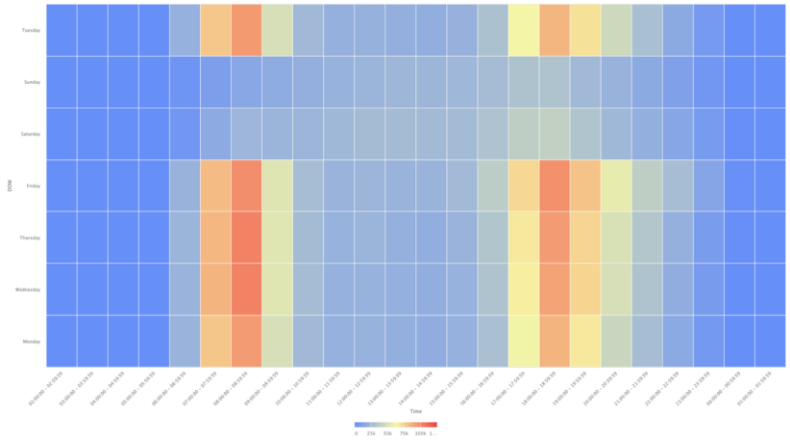

Fig. 2. Heatmap represent journey volume on each hour and day of week

2. Highlight behavioral peaktime compare to difined peaktime from historical period. In the present time, peak time or peak hours period was defined by specific predefine time. However, there still a question whether the pre-defined peak usage period is correctly defined or not compare to the historical data. Fig.3 the author using median value to smooth the visualization[4] which reveals the peak-time from the history data as 06:00 – 10:00 in the morning and 16:00 – 21:00 in the evening which the pre-specified still in the range but not cover all peak-hour period as data given.

Fig. 3. Median value of accumulate journey on each period of time

3. Knowledge discovery on interested station from the journey data. Visualization of transformed data has shown an essential insight of both highest traffic origin and destination station. Fig.4 shows the certain amount of 3 station which most of the passenger travelled in a single journey.

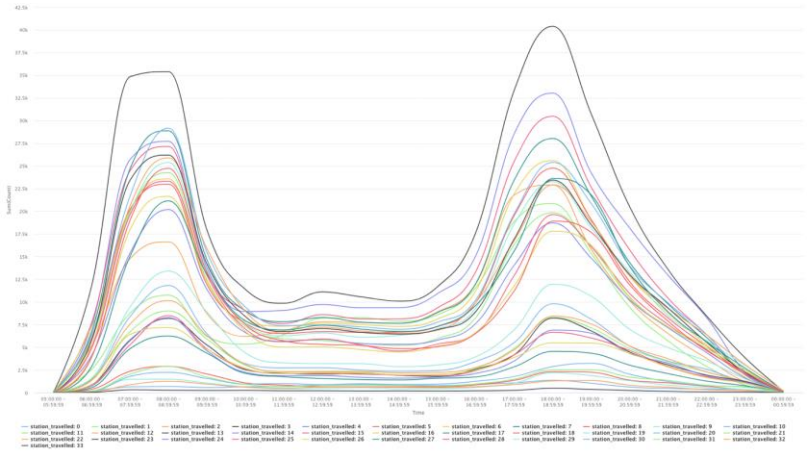

Fig. 4. Line chart represent accumulate count of journey with specific amount of station each passenger travelled in timeframe

4 Discussion

This independent study determines to use a data science knowledge together with domain expert in consumer behavior of the author and advancement of visualization technique to display an insight of the data. We then present the result and conclusion to data owner team who gave us a requirement, consist of one executive level responsible for ticketing and commercial planning department with other two operation level personnel in the same department. After we brought the result to the data owner for a feedback, the suggestion mainly focuses on Geographical integration issue – which the executive level of the data owner suggests that it would be very useful if the data will be allowed to use without blinding and plot on the geography map with other relevant data such as habitat density and public gathering place for further analysis in the future. We expect that in the near future we can use more advance technique and increase more variety sources into the analysis for the public benefit. Moreover, we expect to initiate prediction model to predict the future usage volume of each station in advance.

5 Conclusion

In this digital era, big data is an important key to help develop understanding and to harness a variety and volume of data which will help unlock the better urban and city planning. With this new and advanced in technology it will help create the resilience city development using such knowledge discovered. In the future, we expect to include a variety sources of data to create more challenges which may resulted to disclose the insight behavior of citizen and leads to a better well-being. The study reveals some important point as followed:

- Peak hour mostly appears on the weekday period and usage of the mass transit congested mainly on weekday. Public holiday has a significant impact of plummet of passenger no matter weekday or weekend.
- Habitat density and also center of business district (CBD) can be imply using the journey pattern in the morning and evening period.
- The majority of the passenger will travel only 3 stations which can be determine as the optimization distance between habitat and workplace.

However, due to privacy and national security concern, we have to blind the geographical information in the data which eliminate some vital analysis features.

References

1. Zheng, Y., Trajectory Data Mining: An Overview. *ACM Trans. Intell. Syst. Technol.6, 3, Article 29*, 41 (2015).
2. CBRE., *MRT extensions to open up new opportunities*. Retrieved from Bangkok Post: https://www.bangkokpost.com/business/1702200/mrt-extensions-to-open-up-new-opportunities, last accessed 2019/06.
3. Wirth, R. & Hipp, Jochen., CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (2000).
4. Rong, Kexin, and Peter Bailis. "ASAP: prioritizing attention via time series smoothing." arXiv preprint arXiv:1703.00983 (2017).