Extraction and Visual Analysis of Negative Traffic Events from Weibo Data

Author: Chenyu Zuo

Supervisor: Dr.-Ing. Linfang Ding

Second Supervisor: Prof. Mag.rer.nat. Dr.rer.nat. Georg Gartner

Outline

1. Introduction
2. Related Work
3. Methodology
4. Experiments
5. Conclusion
6. Outlook

1 Introduction

1.1 Background

1.1 Background

Studies about Social Media

- Real time event detection
- Market prediction
- Industry competitive analysis
- Traffic condition information extraction
- Real-time road traffic congestion detection

1.2 Objectives

1. Design text mining methods for negative traffic related information extraction from Weibo text.

2. Derive high-level negative traffic events by using spatial clustering and aggregation methods.

3. Propose visual analytic methods for the exploration of the semantic, temporal and spatial patterns and the correlations among different types of negative traffic events.

2 Related Work

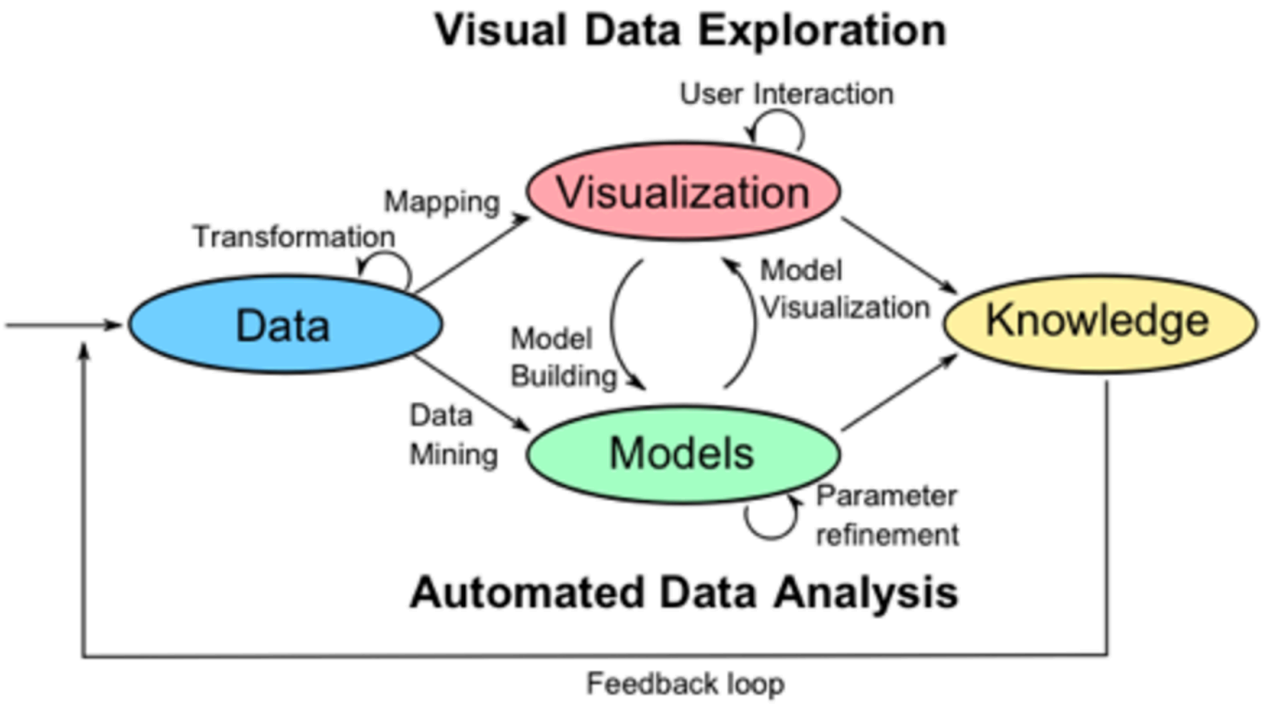
2.1 Text Mining Methods

- Sentiments and opinion analysis
- Natural language processing techniques
- Pre-selected lexicon to extract specific topics

2.2 Clustering Methods

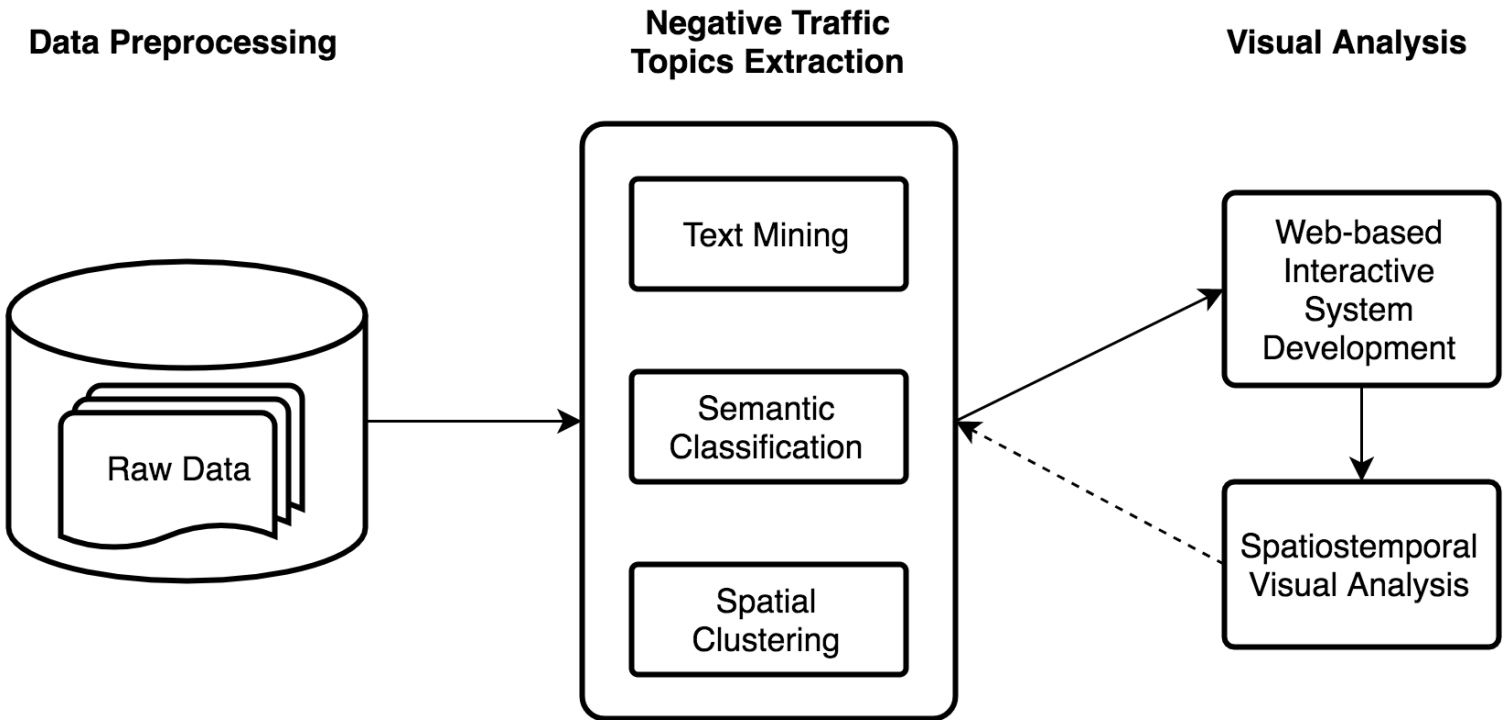
Partitioning methods	Hierarchical methods	Density-based methods	Grid-based methods
Knowledge of resulting number of clusters	Hierarchical decomposition	Detects noise points	Quantizes the data into cells
Assigns all points into clusters	Compact cluster	noise insensitive	calculation time is independent of data dimensions
Spherical-shaped cluster	Spherical-shaped cluster	clusters of arbitrary shape	grid structures
K-means, ISODATA	CURE, BIRCH	DBSCAN	STING

2.3 Visual Analytics



3 Methodology

3.1 Workflow


Data Preprocessing
Negative Traffic Topics Extraction
Visual Analysis
Raw Data
Text Mining
Semantic Classification
Spatial Clustering
Web-based Interactive System Development
Spatiostemporal Visual Analysis

3.1 Data Pre-processing

Raw Data Parameters

Categories	Parameters
Time	idNearByTimeLine, createdAT, createdATUnixTime
Message	msgID, msgmid, msgtext, msgin_reply_to_status_id, msgin_reply_to_user_id, msgin_reply_to_screen_name, msgfavorited, msgsource
Location	geoTYPE, distance, Latitude, Longitude, NAME_1, NAME_2, NAME_3
User	userID, userscreen_name, userprovince, usercity, userlocation, userdescription, userfollowers_count, userfriends_count, userstatuses_count, userfavourites_count, usercreated_at, usergeo_enabled, userverified, userbi_followers_count, userlang, userclient_mblogid

[a] Data exemple: 95, 2014-07-17 04:28:57, 1405571337, 3154521583132051, 3154521583132051, Morning! http://t.cn/RPzg4sq, 0, 0, , 0, , Point, 8200, 31.192111595703125, 121.68149358007813, Shanghai, Shanghai, Pudong, 3963897579, muamuamua, 31, 15, Shanghai , , 18, 95, 70, 0, 2013-12-30 14:22:56, 1, 0, 3, zh-cn, , .

PS: Texts from Weibo data are mostly Chinese.

3.1 Data Pre-processing

Selected Fields

Field Name	Field Value
createdAT	2014-07-17 04:28:57
Latitude	31.192111595703125
Longitude	121.68149358007813
msgtext	Morning! http://t.cn/RPzg4sq

3.1 Data Pre-processing

- **Integrity Check**

Remove records lack of **coordinates**, **time** records and/or **text** content.

- **Deduplication**

Remove redundant records, such some repeating/similar Weibos.

3.1 Data Pre-processing

Text Cleaning

Items	Item Value
Link	http://t.cn/RPzg4sq
Emoji	⌣̈
User Tagging	@ Jack
Punctuation	! , 。 ~ 《》: () 【】
White Space	" "
Digits	345
Tags	#New Year#

3.2 Event Extraction

What to extract?

- Negative Traffic Events （NTE）
- Negative Traffic Topics （NTT）

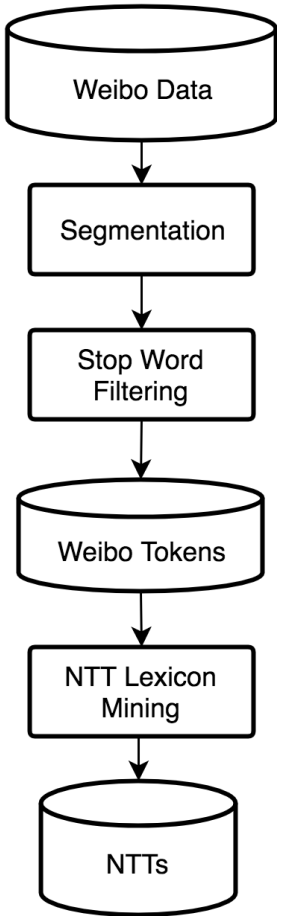
3.2 Event Extraction

Negative Traffic Topics Examples:

- The cars are not moving. There is an accident again.
- Too crowded in the bus today.
- It's so hard to find a taxi here.
- So many people are waiting for the metro.
- No buses are coming.
- One unlucky guy had car crash.
- I have been waiting half an hour for the cab.
- It's so crowded in metro every day when I'm going to work.

3.2 Event Extraction

Wordflow of Event Extraction



3.2.1 Text Segmentation

Example

Raw Text: 延安路上发生了一起车祸。
(An car accident happened at Yan'an Road.)

Segmented: 延安路 发生 车祸
(Yan'an road, happened, accident)

3.2.1 Text Segmentation

Term Frequency – Inverse Document Frequency

$$ittf(t_i) = log\frac{\sum_j dl_j}{\sum_j tf_{i,j}}$$

- t_i - term
- $ittf(t_i)$ - the frequency of term
- dl_j - the length of the document
- $tf_{i,j}$ - t_i in document d_j

Python package: Jieba https://github.com/fxsjy/jieba

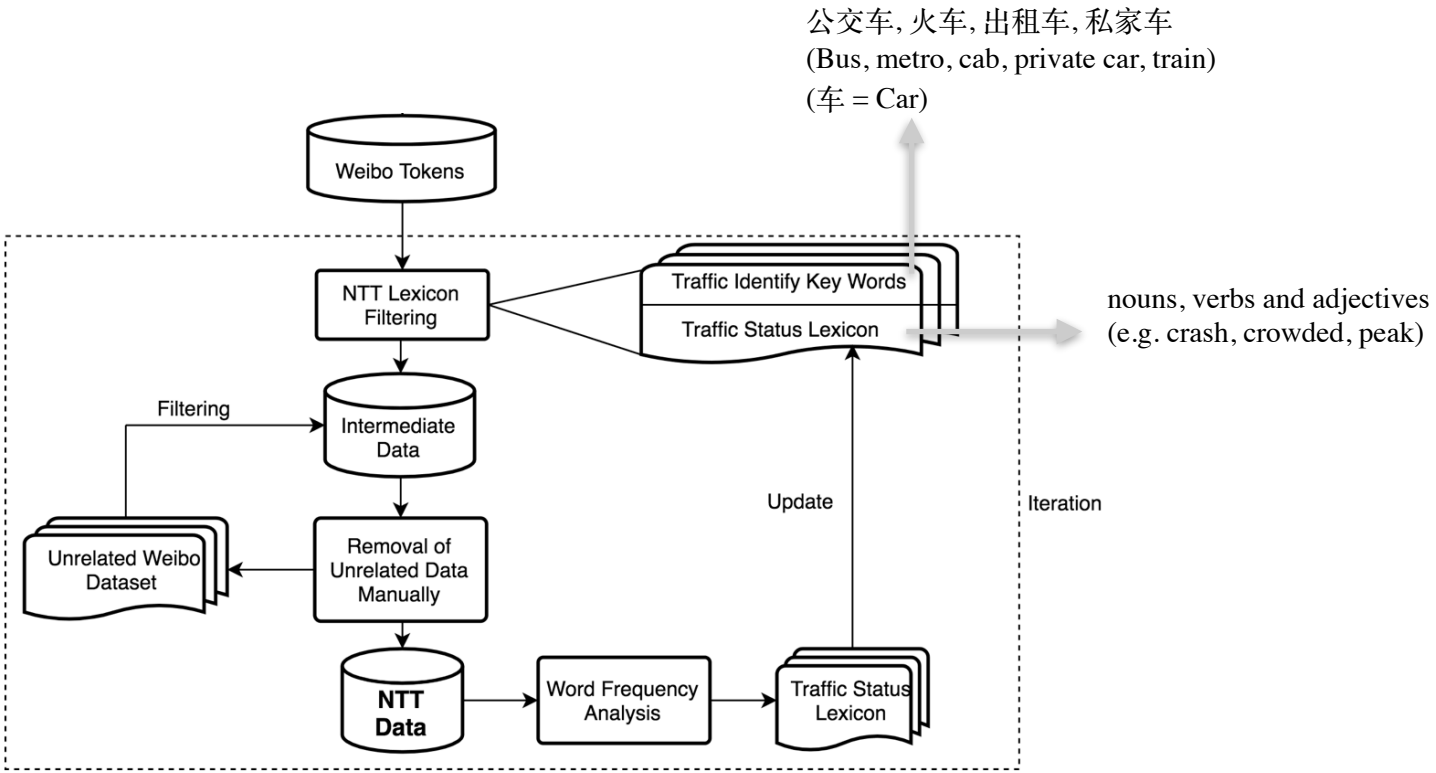
3.2.2 Stop Words Filtering

Stop Words Examples

Languages	Stop Words Examples
Chinese	嗯，哦，后来，还有，或者，我，我们，你的，上述，主要，其次，到底，因为，基本上，亲自，一些，一天，依照，即刻，逐渐，非常，按时，每天
Translation	em, huh, then, and, or, me, our, yours, above, mainly, secondly, actually, because, basically, personally, some, one day, according to, this moment, gradually, very, on time, everyday

A online stop word list: 1609 words.
http://blog.csdn.net/shijiebei2009/article/details/39696571

3.2.3 Lexicon-based NTT extraction



3.2.3 Lexicon-based NTT extraction

Final Traffic Status Lexicon

Categories	Chinese Lexicon	Translation
No taxis	叫不到车，打不到车	cannot find a taxi
	没叫到车，没打到车	have not found a taxi
No cars	没车	no car/bus/taxi
	车呢，车在哪	where is the car/bus/taxi
Congestions	堵	(car) blocking
	挤	crowded
Accident	车祸	car accident
	撞车	car crash
Peak	高峰	peak/rash hour

3.3 Spatial Clustering

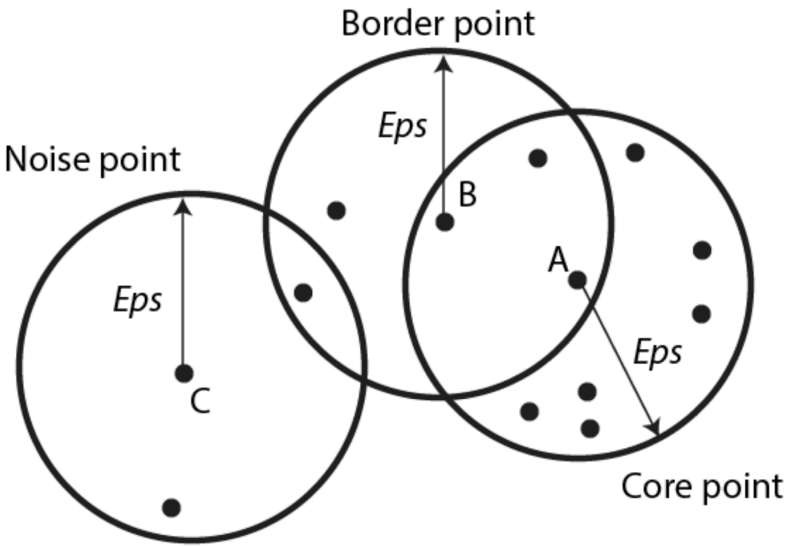
Grid-based Clustering

- Grid-structured
- Quick
- Data dimensions independent
- An overview of a statistical spatial pattern

3.3 Spatial Clustering

DBSCAN Clustering

Eps	*MinPts*	Scale
3.5 km	18	Town
3.5 km	23	Town
350 m	4	Street



3.4 Visualisation

Methods

- Scatter plot map
- Heat map
- Chord diagram
- Alluvial diagram

3.4 Visualisation

Interactivity in the System

- Zoom in/out
- Mouse hovering
- Animation
- Time slider

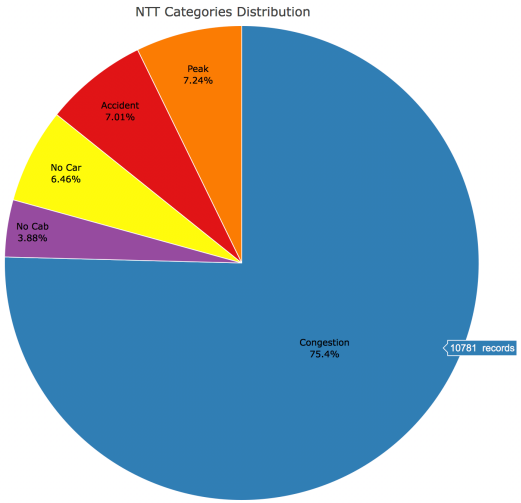
4 Experiments

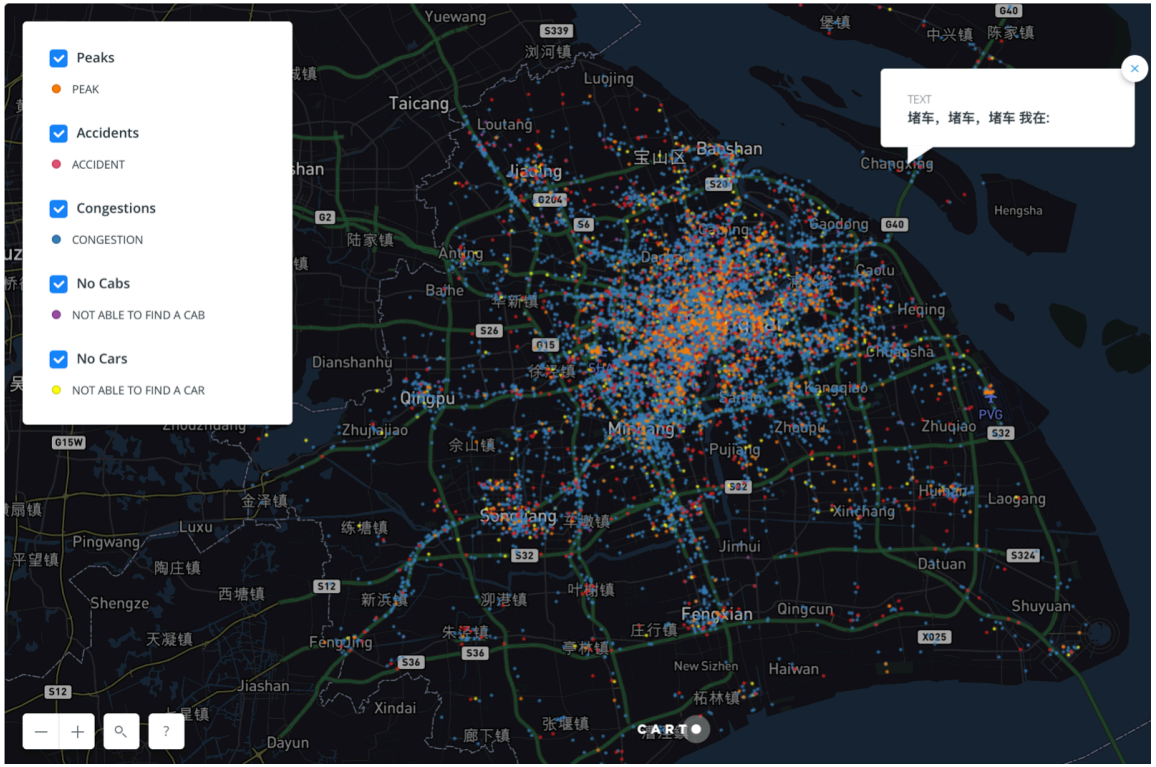
4.1 Test Area Shanghai

Population: 24.15 million in 2016

Area: 6,340 km²

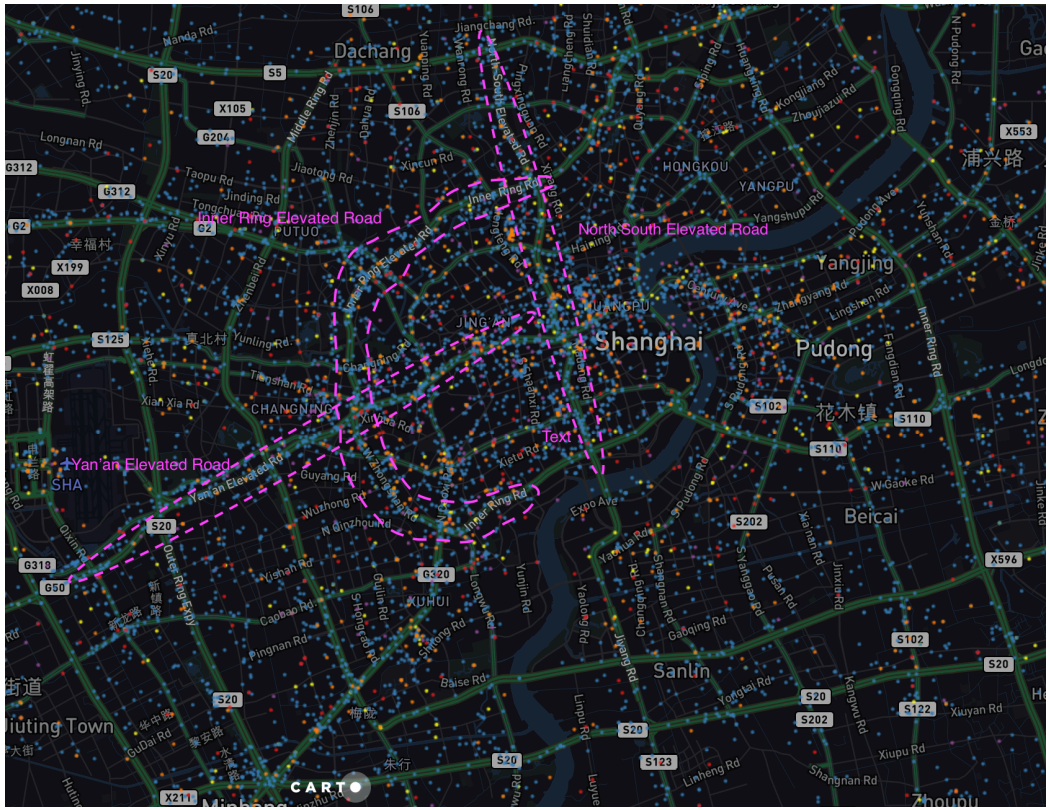
4.1 Overall NTE Patterns





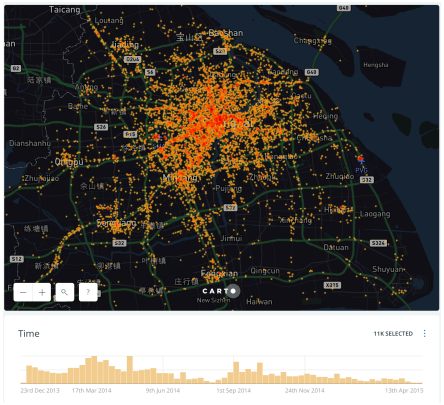
4.2 Overall NTE Patterns

- ACCIDENT
- NO CAB
- NO CAR
- CONGESTION
- PEAK

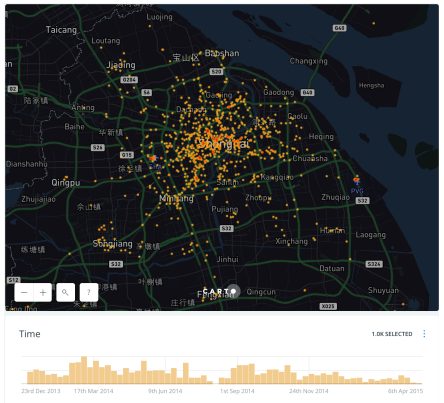


Shanghai Center Area

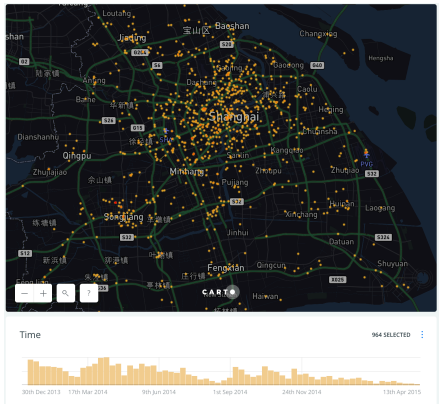
4.3 Semantic NTE Patterns



“Congestion”

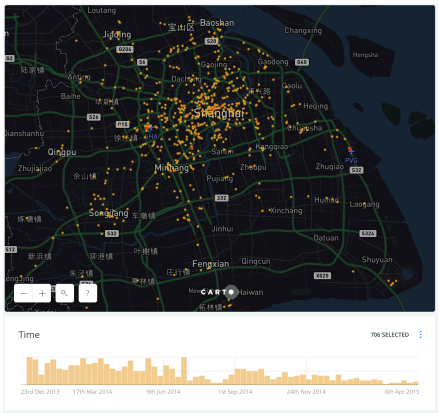


“Peak”

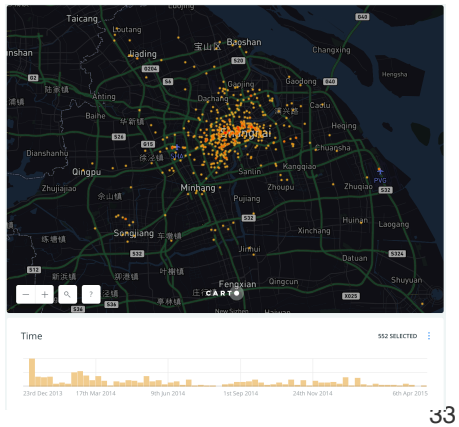


“Accident”

Scatter Plot Maps
with Semantic Categories



“No Car”

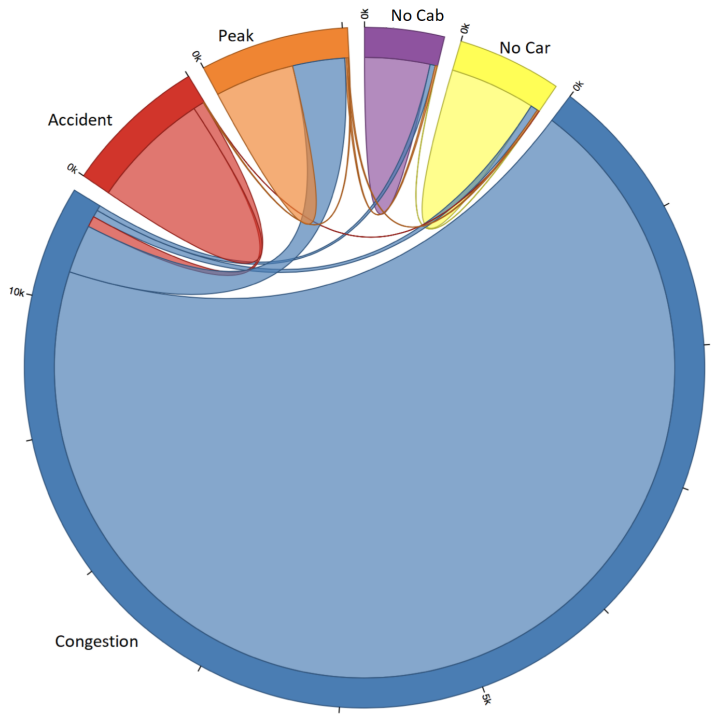


“No Cab”

4.3 Semantic NTE Patterns

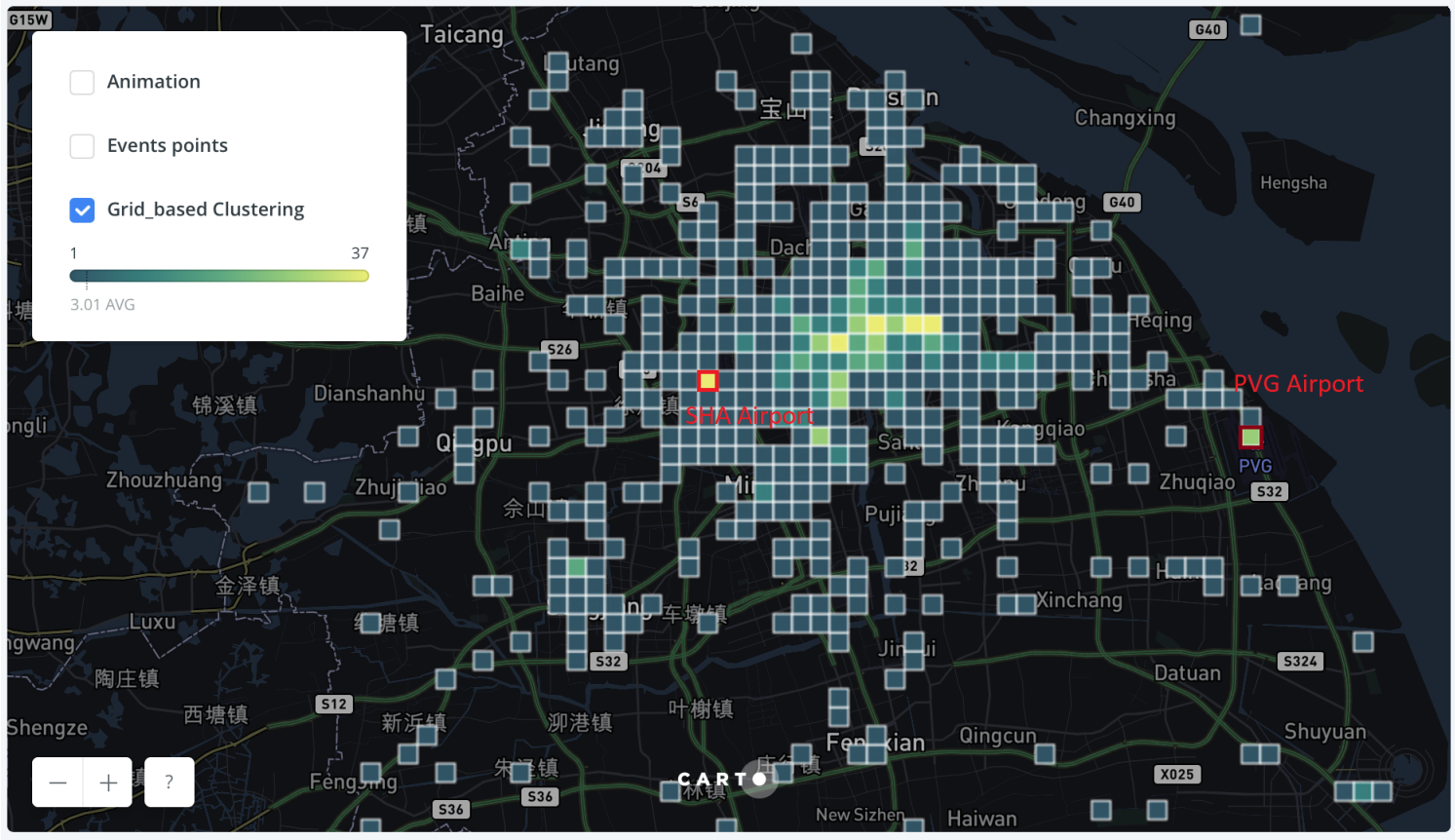
Correlation Matrix

Categories	No Cabs	No Cars	Congestions	Accident	Peak
No Cabs	498	3	40	0	14
No Cars	3	654	55	2	12
Congestions	40	55	10232	83	371
Accident	0	1	83	911	7
Peak	14	12	398	7	605

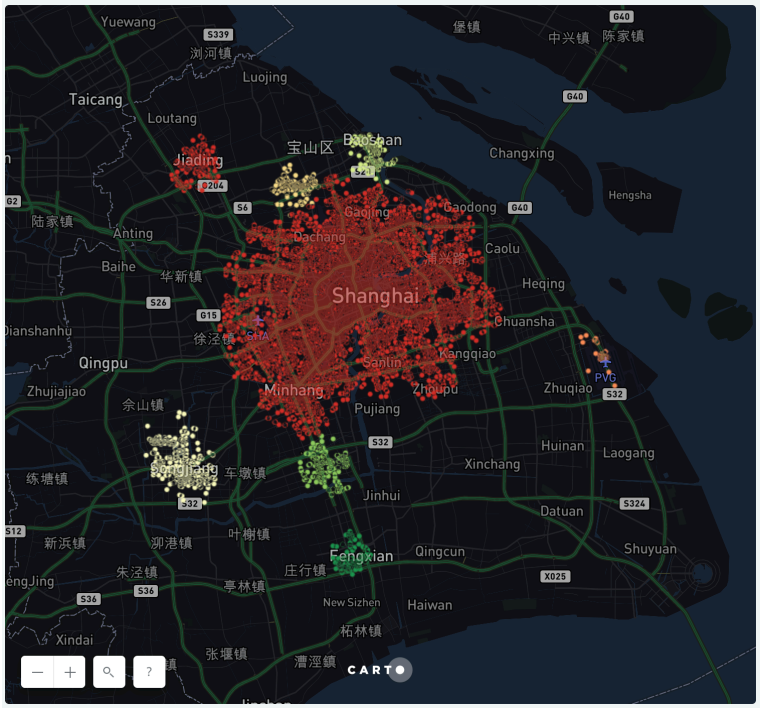


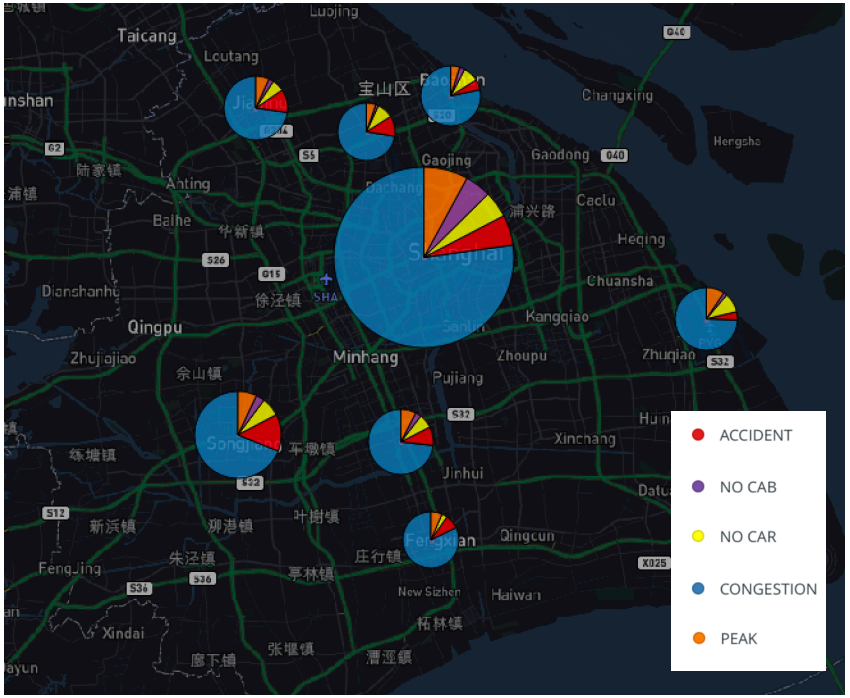
A Chord Diagram of NTEs

4.4 Aggregated NTE Patterns

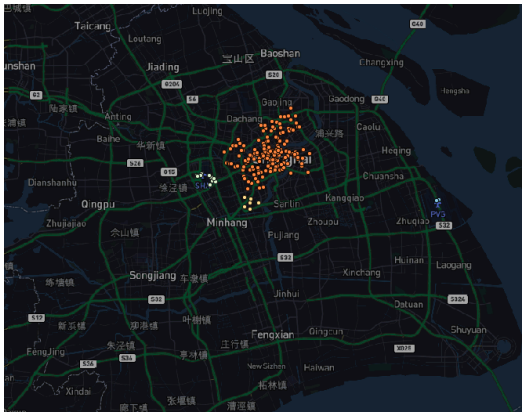


4.4 Aggregated NTE Patterns

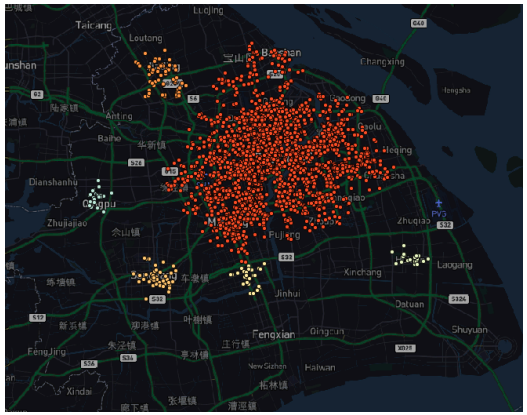




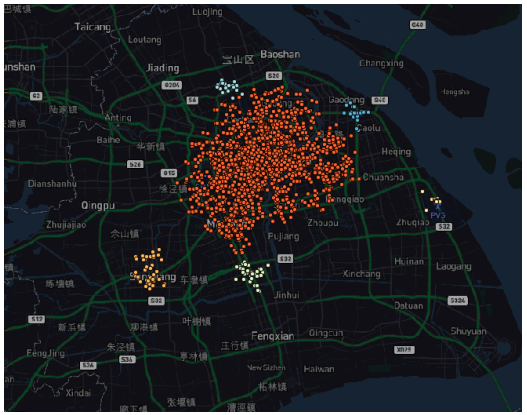
4.4 Aggregated NTE Patterns



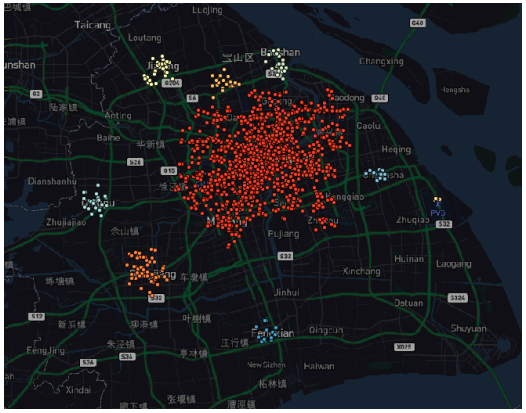
1.00 - 5.00



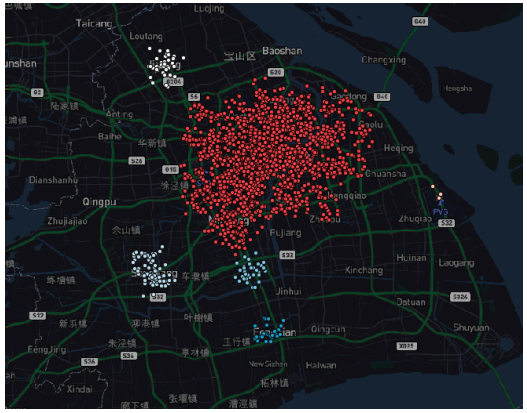
5.00 - 9.00



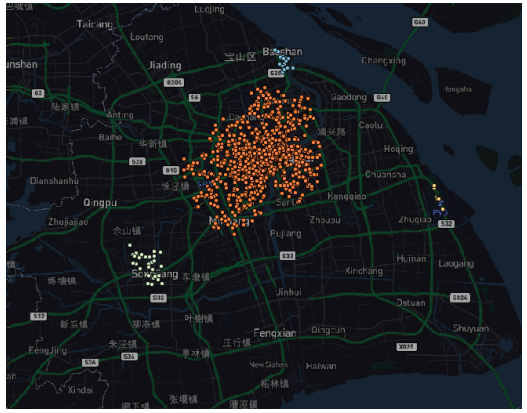
9.00 - 13.00



13.00 - 17.00

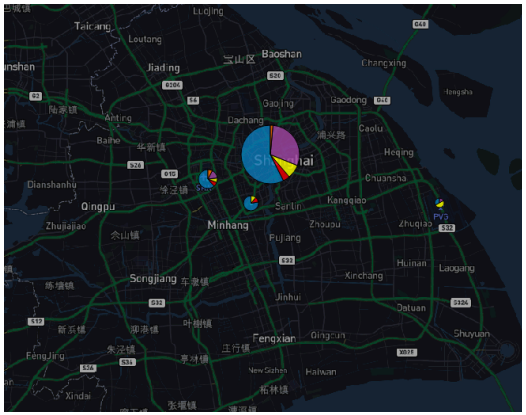


17.00 - 21.00

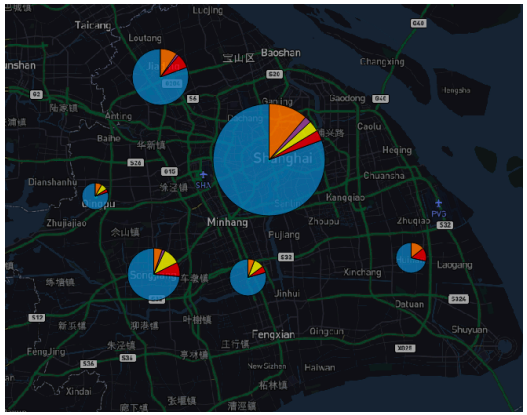


21.00 - 1.00

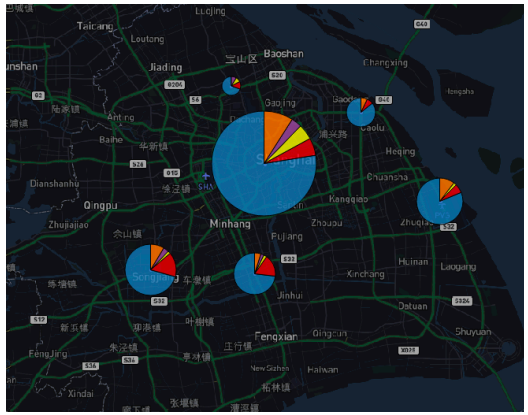
4.4 Aggregated NTE Patterns



1.00 - 5.00

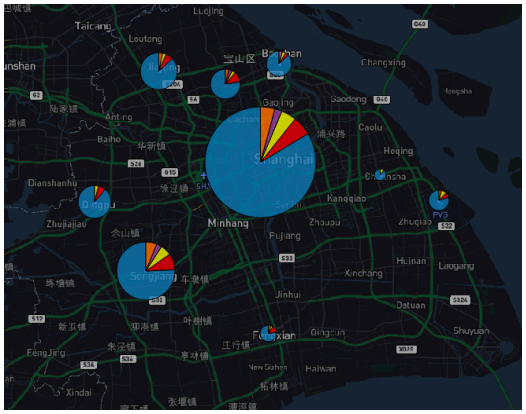


5.00 - 9.00

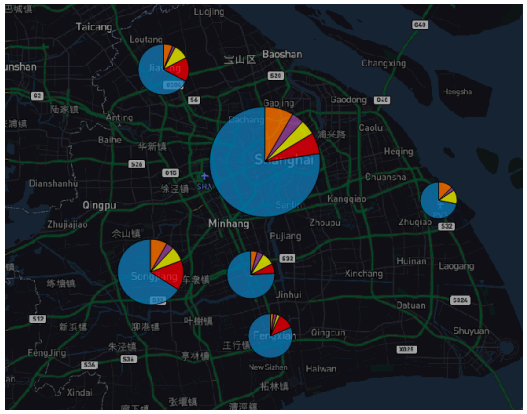


9.00 - 13.00

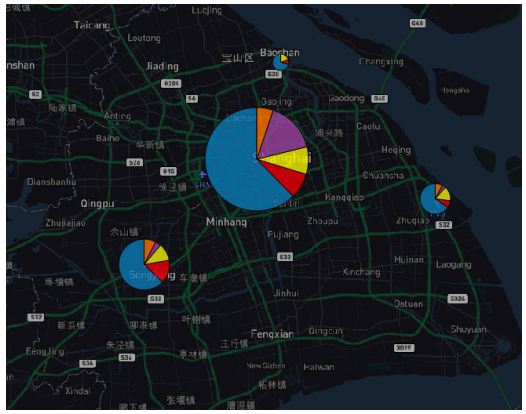
- ACCIDENT
- NO CAB
- NO CAR
- CONGESTION
- PEAK



13.00 - 17.00



17.00 - 21.00



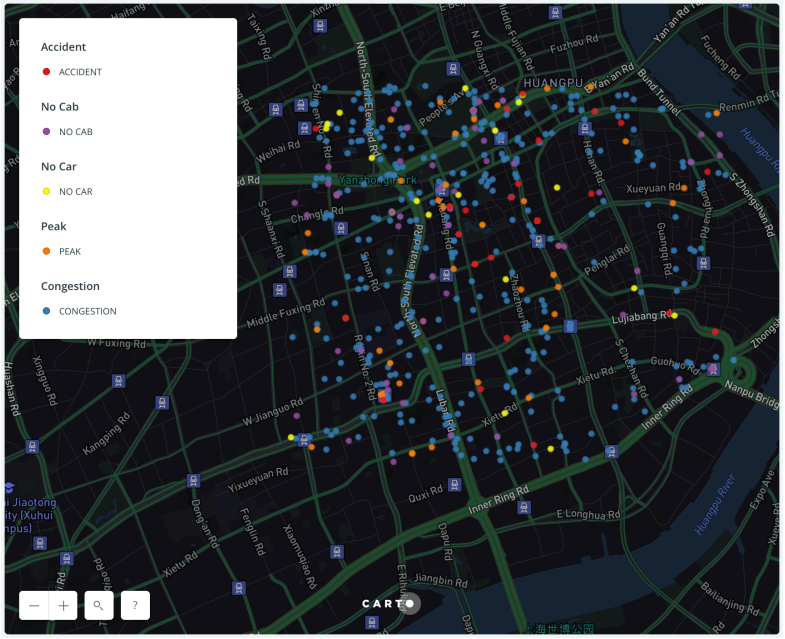
21.00 - 1.00

4.5 Subareas NTE Patterns

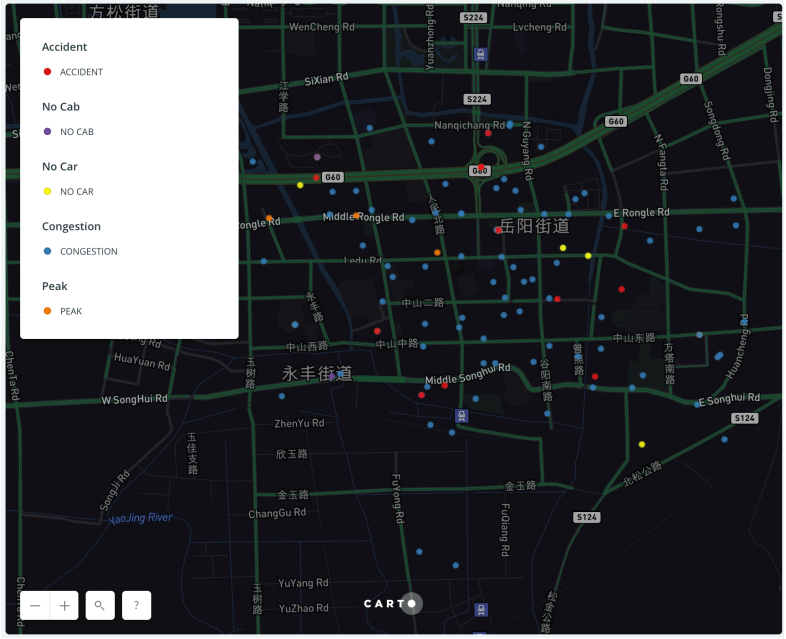
Test Subareas

Area	Latitude Range	Longitude Range
Huangpu District	31.202848 - 31.233495	121.455047 - 121.496953
Songjiang Town	30.986531 - 31.026639	121.204753 - 121.245824

4.5 Subareas NTE Patterns

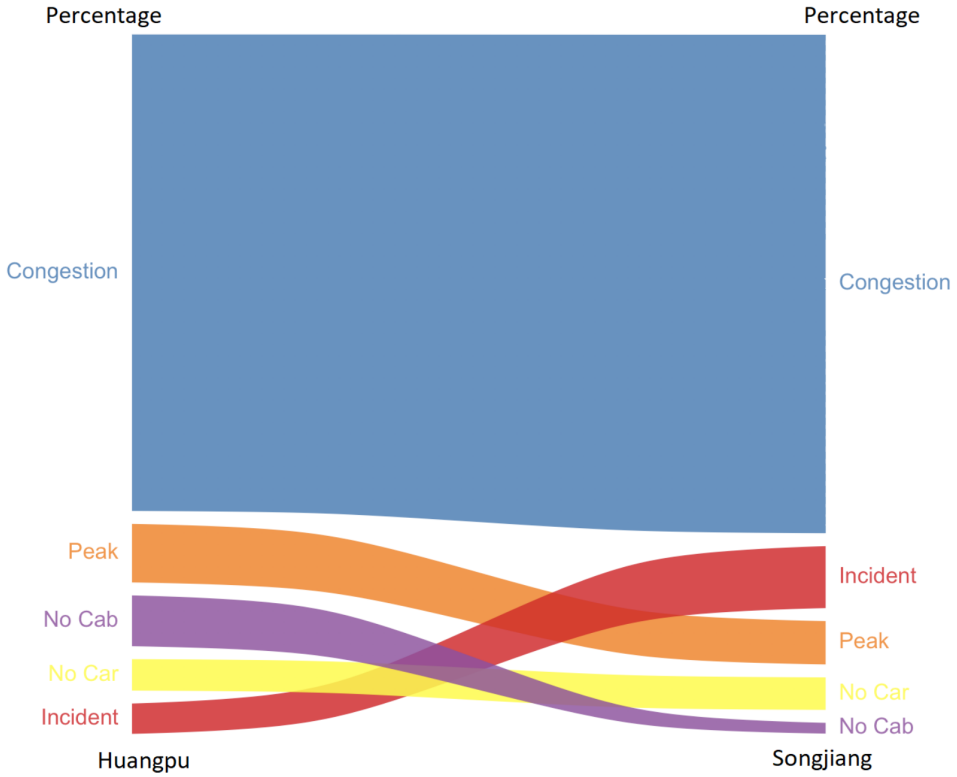


Huangpu District



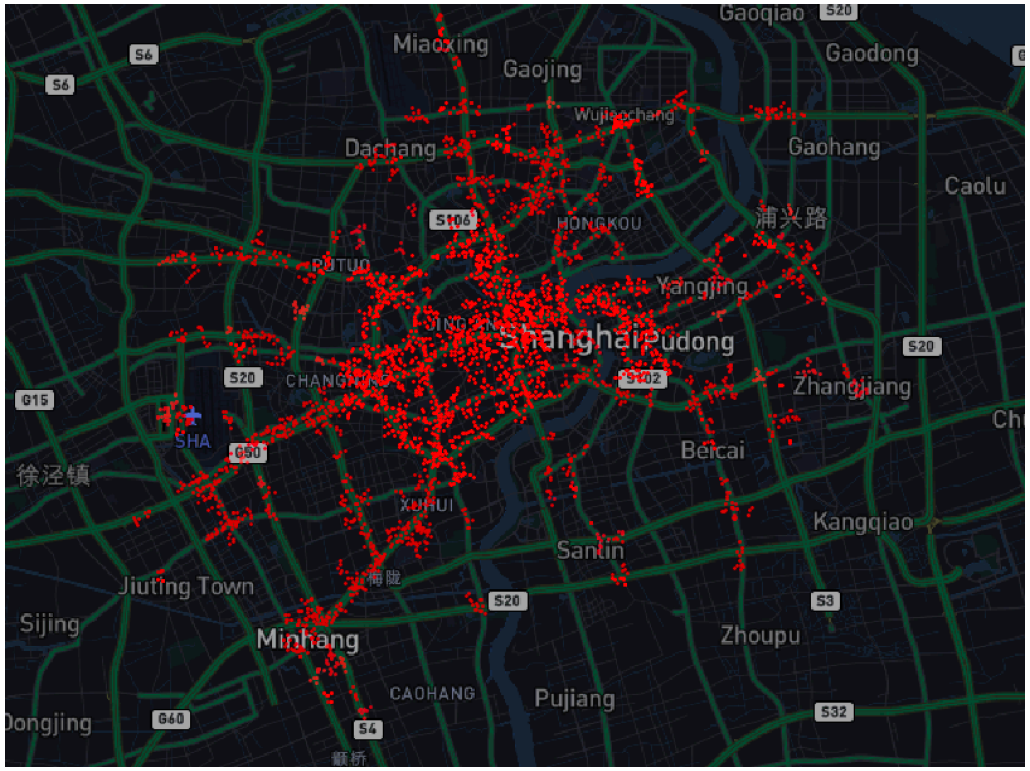
Songjiang Town

4.5 Subareas NTE Patterns



An Alluvial Diagram of Huangpu District and Songjiang Town

4.5 Subareas NTE Patterns



A Scatter Map of NTE Skeleton Extracted by DBSCAN in city centre

5 Conclusion

Conclusion

1. The NTEs information was obtained based on the text mining from social media data.
2. A high-level negative traffic event information is driven by clustering methods.
3. The scatter plot map, the heatmap, the chord diagram and the alluvial diagram were proposed to visualize the NTE patterns.

6 Outlook

Future Work

- Enhancing the temporal analysis (weekday/weekend).
- Adding auxiliary data for analysis (e.g. weather data).
- Analyzing the emotion score according to NTE.
- Integrating more interactive methods in the system (3D).