Extraction and Visual Analysis of Negative Traffic Events from Weibo Data

Author(s):
Zuo, Chenyu; Ding, Linfang; Gartner, Georg; Jendryke, Michael

Publication Date:
2018-01-15

Permanent Link:
https://doi.org/10.3929/ethz-b-000225604 →

Extraction and Visual Analysis of Negative Traffic Events from Weibo Data

Chenyu Zuo*, Linfang Ding*, **, Georg Gartner***, Michael Jendryke****

* Chair of Cartography, TU Munich, Arcisstr. 21, 80333 Munich, Germany. chenyu.zuo@tum.de, linfang.ding@tum.de
** KRDB Research Centre for Knowledge and Data, Faculty of Computer Science, Free University of Bozen-Bolzano, piazza Domenicani 3, 39100 Bozen-Bolzano, Italy.
*** Research Group of Cartography, TU Vienna, Favoritenstr. 9, 1040 Vienna, Austria. georg.gartner@tuwien.ac.at
**** LIESMARS, Wuhan University, Luoyu Road 129, Wuhan, China. michael@jendryke.de

Abstract. Social media has an increasingly significant importance in people's daily life during past few years. Social media data has been widely studied in a variety of disciplines for different applications, e.g., crisis management and urban planning. In this work, we focus on the extraction and analysis of traffic related events, especially negative traffic events(NTE), e.g. congestion, car accidents, from social media data. Firstly, we identify the terms related to NTEs. Secondly, based on those terms, an iterative lexicon-based text mining technique is applied to extract NTEs from social media data. Thirdly, we calculate the statistics and visualize the spatiotemporal patterns of the NTEs. A web-based interactive visualization system is developed for visual analysis of NTEs. We use one year Sina Weibo data in Shanghai as our test data and present our preliminary results.

Keywords. Social Media, Negative Traffic Events, Event Extraction, Visual Analytics, Spatiotemporal Mining, Text Mining

1. Introduction

Social media has become part of people's daily life and changed the spread of information (Seger, 2011; Tuten & Solomon, 2014). A wide range of topics, e.g. food, work, traveling, sports, entertainment and emotion are posted by a large amount of users. Social media data has been studied and applied in various disciplines, e.g., event detection (Sakaki, Okazaki, & Matsuo, 2010), market prediction (Bollen, Mao, & Zeng, 2011), sentiment analysis (Agarwal, Xie, Vovsha, et al., 2011).

Recently, social media data is also widely used as an information source for traffic condition detection, such as traffic congestion, accidents and road works (Wanichayapong, Pruthipunyaskul, Pattara-Atikom, & Chaovalit, 2011; McHugh, 2015). A *negative traffic event* can be defined as a real-world traffic-related occurrence that had a negative effect on people's traveling, i.e. car congestion, car accident, crowded metro/bus in peak hours and long waiting time for public transports. Detection and analysis of NTEs from social media data could improve traffic management and car navigation.

In this study we focus on the detection and visual analysis of NTEs from social media data. Section 2 outlines methodology in NTE detection and visual analysis from social media data. Section 3 introduces the test data, the interactive visual system and results. Section 4 concludes the study and proposes further research ideas of possibilities for NTE analysis.

2. Methodology

In this section, we will introduce the methodology for NTEs extraction and visual exploration, which mainly consists of data preprocessing, text mining, spatial clustering, aggregation and visualization. Figure 1 shows the workflow of NTE extraction and visual analysis from social media data.

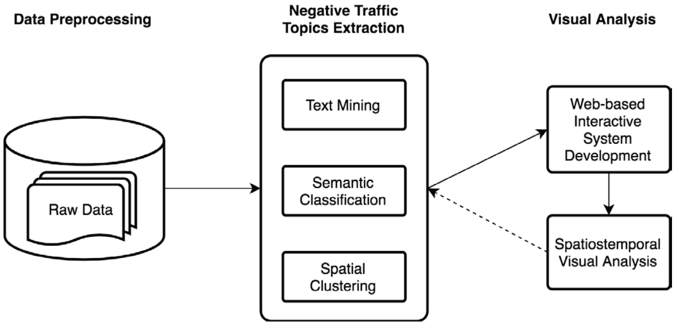

Data Preprocessing
Negative Traffic Topics Extraction
Visual Analysis
Raw Data
Text Mining
Semantic Classification
Spatial Clustering
Web-based Interactive System Development
Spatiostemporal Visual Analysis

Figure 1. The workflow for NTEs extraction and visual analysis.

2.1. Negative Traffic Event Extraction and Clustering

2.1.1 Pre-processing

The raw social media data typically includes 36 parameters. Before data processing, data parameter selection and data cleaning are important steps.

To reduce the amount of the data volume, only relevant parameters are selected for the analysis. As this work focuses on NTE extraction, time *record*, *location* and *texts* are the vital parameters. Therefore, those parameters are selected as inputs for the processing framework. Besides, raw data is often incomplete, inconsistent or contains abnormal values. Therefore, data cleaning is performed to remove those records. Also, in each text, some contents are less or not helpful for event extraction, such as special characters, links, hashtags and punctuations. Those information is also removed during the pre-processing.

2.1.2 Event Extraction based on Text Mining Method

To extract negative traffic event, a text mining method is designed and applied to the preprocessed data. The detail of the method is described in the following:

a) *Tokenization* is generally the necessary first step in text processing (Xue et al., 2003). It segments character streams into meaningful terms (tokens). *TF-IDF (term frequency–inverse document frequency)* (Forman, 2008) is the most widely used machine learning method for text mining. Jieba[1] is a Python package used for text tokenization in this work. It is based on a prefix dictionary structure to achieve efficient word graph scanning, build a directed acyclic graph for all possible word combinations. Following this, the most probable combinations are identified based on the TF-IDF. For unknown words, a *Hidden Markov Model* is used with the Viterbi algorithm.

b) *Stop word filtering* is to eliminate *stop words*, which are unlikely to assist further semantic understanding of computer algorithms, and those words also appear frequently in the text (Fox, 1989). The stop word list[2] used in this work includes 1609 words.

c) *NTE identification* is to extract NTE-related texts. We assume that NTE-related texts must contain two features: *traffic objects (tObj)* and *traffic status (tSta)*. Specifically, *tObj* includes vehicle names, such as car, bus, metro and cab. *tSta* includes nouns, verbs and adjectives commonly used in NTE describing, such as crash and crowded. Firstly, NTE-related candidates are identified as records contains both *tObj* and *tSta* tokens. Secondly, a manual check is per-

[1] https://github.com/fxsjy/jieba

[2] http://blog.csdn.net/shijiebei2009/article/details/39696571

formed to remove NTE-unrelated records from candidates. After each round of *NTE identification*, *tSta* is updated based on token frequency analysis. The *NTE identification* processing is iterative with the updated *tSta*.

d) *Semantic classification of NTEs* is to classify NTEs by their semantic meaning. As the NTE-status are related to *tSta*, NTEs are classified according to *tSta* classification. *tSta* tokens are classified into different semantic groups, such as congestion, accident, peak hours.

2.1.3 Clustering and Aggregation

Clustering and aggregation of NTEs refer to the group of the similar NTEs that helped to access the importance and impact of certain topics. Grid-based clustering and DBSCAN clustering methods are applied to drive high-level spatial patterns, because those two clustering methods do not require a pre-knowledge of spatial distribution of data. Grid-based clustering and statistical aggregation could generate clustering patterns quickly (Grabusts & Borisov, 2002). With DBSCAN, NTEs at different scales of arbitrary patterns could be generated, such as district scale and street scale (Ester, Kriegel, Sander, et al., 1996). Besides, NTEs could also be categorized by different time periods of a day. By visualizing NTEs according to different time periods, temporal distribution of NTEs could be explored and analyzed.

2.2. Visualization

Visualization provides a powerful mean of data understanding. By mapping data attributes to visual properties such as position, size, shape, and color, visualization designers leverage perceptual skills to help users discern and interpret hidden patterns (Card, Mackinlay, & Shneiderman, 1999).

In this work, to allow the visual exploration of the large amount of extracted NTE patterns for domain experts (e.g. traffic investigator), we design and develop an interactive web-based system with spatial and temporal visualization components, including a map view, a histogram view with a time slider, and a statistical view showing information visualizations like chord diagrams. We apply scatter plot maps, temporal histograms, heatmaps and chord diagrams to reveal the semantic, temporal and spatial NTE patterns and the correlations among different NTE categories. A scatter plot map is to show datasets on maps using dots. It is simple and accurate to represent the spatial distribution of data (Anselin, Syabri, & Kho, 2006). A heatmap is a graphical representation of data where the individual values are represented as colors. A chord diagram could show the correlation among categories. We can answer some important questions by exploring the NTEs in this system, for instance: Where are NTE hotspots? Which area contains most *Congestion*? When is the most busy time for traffic?

3. Experiments

3.1. Test Data

In this study we use Sina Weibo (WB) data as our test dataset. WB is a Chinese microblogging website, with over 297 million active users per month until 2016 (Center, 2016). The dataset has all the records in Shanghai in 2014, in total 11590886 records. It was saved in the form of CSV (Comma-separated values) file. Data description is shown in Table 1. The original data contains four fields of information: time, location, massage and user.

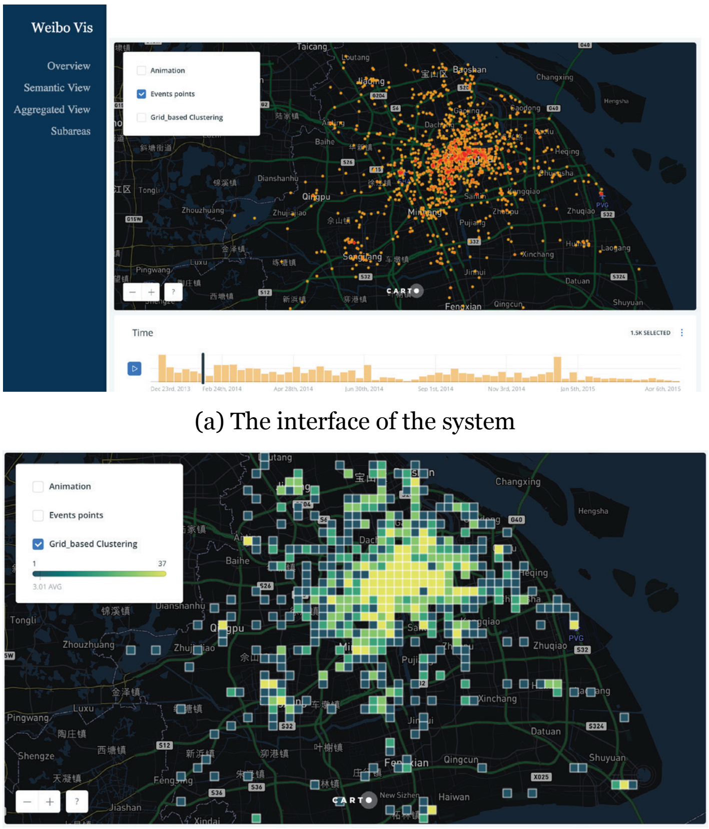
Categories	**Parameters**
Time	idNearByTimeLine, createdAT, createdATUnixTime,
Location	geoTYPE, distance, Latitude, Longitude, NAME_1, NAME_2, NAME_3
Message	msgID, msgmid, msgtext, msgin_reply_to_status_id, msgin_reply_to_user_id, msgin_ reply_to_screen_name, msgfavorited, msgsource
User	userID, userscreen_name, userprovince, usercity, userlocation, userdescription, userfollowers_count, userfriends_count, userstatuses_count, userfavourites_count, usercreated_at, usergeo_enabled, userverified, userbi_followers_count, userlang, userclient_mblogid
Data example: 95, 2014-07-17 04:28:57, 1405571337, 31.192111595703125, 121.68149358007813, Shanghai, Shanghai, Pudong 3154521583132051, 3154521583132051, Morning! http://t.cn/RPzg4sq, 0, 0, , 0, , Point, 8200, 3963897579, muamuamua, 31, 15, Shanghai , , 18, 95, 70, 0, 2013-12-30 14:22:56, 1, 0, 3, zh-cn, ,	

Table 1. Data Description

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

Table 2. A Grey Shaded Matrix

We apply the NET extraction methods proposed in Section 2 to our dataset and get 14297 extracted NTEs. In total we have five categories of our NTEs, namely *Congestion, Accident, Peak, No car* and *No Cab*. Table 2 shows the confusion matrix of extracted NTEs. The cell background colors stand for the numbers of NTEs, and the darker the cell is, the larger the number is. *Congestion* takes the largest part of NTEs. Interestingly, a large number of *Peak* records do also belong to *Congestion*.



☐ Animation
☑ Events points
☐ Grid_based Clustering

(a) The interface of the system

☐ Animation
☐ Events points
☑ Grid_based Clustering

(b) A heatmap of NTEs

Figure 2. The interactive visual analysis system.

3.2. Visual Analysis Results

Figure 2(a) shows the graphic user interface of "*WeiboVis*". The left part is a navigation panel, users can choose map categories from the navigator. The right part of the interface is the mapping panel which shows the corresponding visualizations according to users' settings. Figure 2(b) shows the NTE heatmap in entire Shanghai. From the map we can see that NTEs distributed radically. There is one big cluster located in the city center, some smaller clusters located in the outskirt areas. Generally, the density of the NTEs are distributed radically. The peak grid is located in city center, including 37 NTEs. The average NTE number of all the grids is 3.01.

To show the correlations among different semantic categories intuitively, we visualize it with a chord diagram shown in Figure 3. The overview of the number of NTEs in each category as well as the correlation between pairs of categories can be easily observed. Interestingly, nearly half of *Peak* events are also associated with *Congestion*.

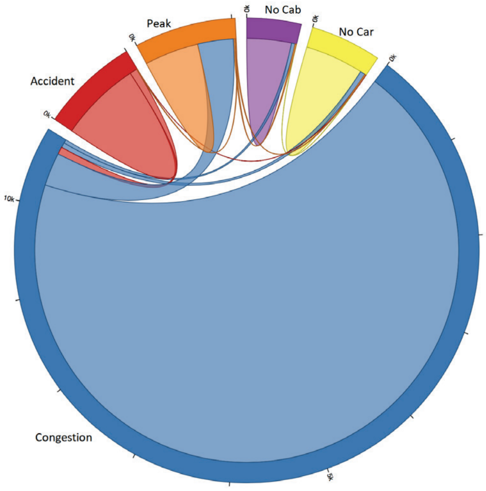


Figure 3. A chord diagram view of correlation among the NTEs semantic categories.

4. Conclusion and Future Work

This study presented a novel approach for NTE extraction from social media data and spatiotemporal analysis. Specifically, a semi-automatic iterative text mining method was designed and performed to extract NTEs from texts. Aggregations were driven based on clustering methods. Visual analytic methods were proposed to explore temporal, spatial and semantic patterns.

In the future, the accuracy of text mining could be improved. Besides, more semantic analysis could be done, for instance, analyzing NTEs with human emotions. Moreover, binding auxiliary data into social media data to analysis more patterns, e.g. with weather data, we can explore the relationship between NTEs and bad weather (i.e. rain, snow and etc.). Beside, usability tests will be done to evaluate the developed web-based visual analytical system.

Acknowledgement

This work is funded by the project "Geospatial Information Services for Smart Cities Driven by Big Data".

References

Agarwal, A., Xie, B., Vovsha, I., et al. (2011). Sentiment analysis of twitter data. In Proceedings of the workshop on languages in social media (pp. 30–38).

Anselin, L., Syabri, I., & Kho, Y. (2006). Geoda: an introduction to spatial data analysis. Geographical analysis, 38(1), 5–22.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1–8.

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). Readings in information visualization: using vision to think. Morgan Kaufmann.

Center, W. D. (2016, 12). Weibo user report 2016. Retrieved from data.weibo.com/report/reportDetail?id=346

Ester, M., Kriegel, H.-P., Sander, J., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, pp. 226–231).

Forman, G. (2008). Bns feature scaling: an improved representation over tf-idf for svm text classification. In Proceedings of the 17th acm conference on information and knowledge management (pp. 263–270).

Fox, C. (1989). A stop list for general text. In Acm sigir forum (Vol. 24, pp. 19–21).

Grabusts, P., & Borisov, A. (2002). Using grid-clustering methods in data classification. In Parallel computing in electrical engineering, 2002. parelec'02. proceedings. International conference on (pp. 425–426).

McHugh, D. (2015). Traffic prediction and analysis using a big data and visualisation approach. Department of Computer Science, Institute of Technology Blanchardstown.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on world wide web (pp. 851–860).

Seger, J. (2011). The new digital [st] age: Barriers to the adoption and adaptation of new technologies to deliver extension programming and how to address them. Journal of Extension, 49(1), n1.

Tuten, T. L., & Solomon, M. R. (2014). Social media marketing. Sage.

Wanichayapong, N., Pruthipunyaskul, W., Pattara-Atikom, W., & Chaovalit, P. (2011). Socialbased traffic information extraction and classification. In Its telecommunications (itst), 2011 11th international conference on (pp. 107–112).

Xue, N., et al. (2003). Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processing, 8(1), 29–48.