Crowd Sensing of Weather Conditions and Traffic Congestion Based on Data Mining in Social Networks

Rita Tse[1], Lu Fan Zhang[1], Philip Lei[1,2,3], and Giovanni Pau[2,3(✉)]

[1] Computing Program, Macao Polytechnic Institute,
Rua de L. Gonzaga Gomes, Macao, China
{ritatse, philiplei}@ipm.edu.mo,
vivianppmonkey@gmail.com
[2] Computer Science Department, University of California, Los Angeles, USA
gpau@cs.ucla.edu
[3] Lip6 – University Pierre et Marie Curie, Paris, France

Abstract. In recent years, the growing prevalence of social networks makes it possible to utilize human users as sensors to inspect city environment and human activities. Consequently, valuable insights can be gained by applying data mining techniques to the data generated through social networks. In this work, a practical approach to combine data mining techniques with statistical analysis is proposed to implement crowd sensing in a smart city. A case study to analyze the relationship between weather conditions and traffic congestion in Beijing based on tweets posted on Sina Weibo platform is presented to demonstrate the proposed approach. Following the steps of raw dataset pre-processing, target dataset processing and statistical data analysis, analytic corpus containing tweets related to different weather conditions, traffic congestion and human outdoor activity is selected to test causal relationships by Granger Causality Test. The mediation analysis is also implemented to verify human outdoor activity as a mediator variable significantly carrying the influence of good weather to traffic congestion. The result demonstrates that outdoor activity serves as a mediator transmitting the effect of good weather on traffic congestion.

Keywords: Smart city · Social networks · Data mining · Weather condition · Traffic congestion · Mediation analysis

1 Introduction

Social networks are becoming increasingly popular in the information era, with the ability to allow people sharing their perspectives upon different areas of urban life. People can communicate with each other and express their own voice through those platforms. Most importantly, the emergence of social networks makes it possible to study human life in a new way due to the fact that the tweets posted on the social networks can reflect people's opinions and emotions [1]. Recently, it has become a new and effective way to research in data mining field by leveraging social media data

O. Gaggi et al. (Eds.): GOODTECHS 2016, LNICST 195, pp. 353–361, 2017.
DOI: 10.1007/978-3-319-61949-1_37

sources in smart cities. These researches involve tackling the challenge of social science, public health, and also weather and traffic conditions at urban management scale in a view of smart city [2–5]. Wang et al. [6] investigated the value of Chinese social media for monitoring air quality trends and the related public perceptions and response. The study verified quantitatively that message volume in Sina Weibo is indicative of true particle pollution levels. Pan et al. [7] described the detected traffic anomaly by mining representative terms from the social media that people posted when the traffic anomaly happened. Also, they demonstrated the effectiveness and efficiency of the method using a dataset of tweets collected from Sina Weibo, a Twitter-like social site. Cool et al. [8] established an approach to use the traffic intensity data originated from minute-by-minute data coming from single inductive loop detectors to analyze the relationship between weather, road safety, traffic speed, and traffic intensity and then investigated the impact of various weather conditions on traffic intensity. Zeng et al. [9] constructed a dynamic evolution network of traffic congestion by the tweets of the online users of social media. The present work aims to propose a simple solution which incorporates statistical analysis into data mining processes to crowd sense traffic conditions in a smart city [10]. The analysis is based on the tweets about weather conditions, traffic congestion and human outdoor activity posted on Sina Weibo platform. The work proceeds to reveal the relationship between weather conditions and traffic congestion in Beijing city of China.

2 Methodology

There are three primary stages in this work: data pre-process stage, data process stage and data analysis stage. Initially, the raw dataset needs to be preliminarily refined in the data pre-process stage to remove noise and to be reformatted prior to further processes. Then the refined dataset can carry on some fundamental mining procedures like word segmentation and frequency analysis to help construct lexicon used in selecting target dataset (analytic corpus). Next, the target dataset is filtered out based on the lexicon and a daily based tweets count file associated with weather conditions, outdoor activity and traffic congestion is generated during the data process stage, which is then served as source file for data analysis. In the last stage, Granger Causality Tests are performed to make sure there are causal relationships between weather and traffic variables. To figure out the indirect relationship between weather conditions and traffic congestion, outdoor activity is introduced as a mediator, and also mediation analysis is conducted to model the relationships among weather conditions, outdoor activity and traffic congestion.

As demonstrated in Fig. 1, data pre-process stage contains two main tasks: data cleansing and lexicon generation. Next, target dataset (analytic corpus) is selected as well as daily based tweets count CSV file is generated in the data process stage. Lastly, Granger Causality Test and mediation analysis are conducted in the data analysis stage.

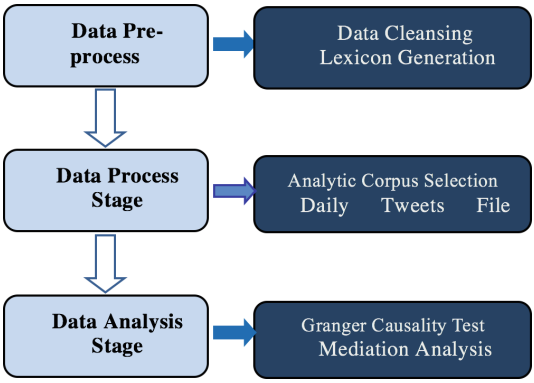

Fig. 1. Data mining major steps.

2.1 Data Pre-processing

To begin with, the data integrity of the raw dataset needs to be checked before any further operations. Since the raw dataset was crawled from the Internet through Sina Weibo API, some redundant tweets may occur during the crawling process. After removing redundant tweets, we used various regular expressions to detect and remove different types of noise data that were not useful in current study. These included picture and video sharing, check-in and empty tweets.

2.1.1 Lexicon Generation

The basis of this work is to develop appropriate lexicon used for filtering out analytic corpus. Unlike sentimental lexicon, it is difficult to find comprehensive and authoritative domain lexicon for weather conditions/outdoor activity/traffic congestion on the Internet. Consequently, the lexicon used in this work is a joint effort of both the initial lexicon of weather conditions, outdoor activity and traffic congestion collected from the Internet and keywords (related to weather conditions, outdoor activity and traffic congestion) with relatively high occurrences selected from the result of word frequency analysis.

Initially, lexicons related to weather conditions, outdoor activity and traffic congestion are collected from Internet which serves as the first part of the lexicon used in this work. Prior to word frequency analysis, content of tweets in the refined dataset are segmented using Rmmseg-cpp, a powerful Chinese word segmentation tool fully implemented by C++. Although there are many word segmentation tools based on different segmentation algorithms available on the Internet, they are mainly designed for English contents. However, most users of Sina Weibo are Chinese which means most tweets in Sina Weibo are written in the Chinese language. Therefore, the Rmmseg-cpp is selected to perform this Chinese word segmentation task in this work due to less memory consumption and fast segmentation speed. Additionally, Rmmseg-cpp is a Ruby gem which can be easily incorporated with Ruby environment and this work is mainly implemented by Ruby. The output of word segments are stored in one TXT file and then a word frequency analysis program is implemented to count

each distinct word occurrence in the file, in which the words are rearranged based on their frequencies to select words related to weather conditions, outdoor activity and traffic congestion with high frequency. In the end, the initial lexicons collected from the Internet are combined with those words selected from word frequency analysis to serve as the ultimate lexicon used in this work.

In the relationship between weather conditions and traffic congestion, there are three main categories: weather conditions, outdoor activity and traffic congestion in the combined dictionary. In order to figure out the direct relationship between weather conditions and traffic congestion, words reflect different weather conditions and traffic congestion are selected. For the weather category, two subcategories are preliminarily divided: good weather and bad weather. Good weather contains words or expressions with respect to normally good weather, like '阳光' (sunshine), '明媚' (radiant and enchanting), '晴朗' (serene), '天气真好' (nice day), '好天气' (lovely weather). On the contrary, bad weather contains words or expressions related to severe weather which may have influences on human outdoor activities, such like '下大雨' (pour), '下大雪' (snow heavily), '寒风' (cold wind). As to the traffic congestion category, words like '堵车' 堵车(traffic jam), '堵塞' (choked), '堵死' (block off) are selected.

On the other hand, the indirect relationship between weather conditions and traffic congestion via outdoor activity is studied. As mentioned before, the outdoor activity serves as the mediated bridge between the weather conditions and traffic congestion where weather conditions have influence on human outdoor activities, and also human outdoor activities affect the traffic congestion. Therefore, for the outdoor activity words which may have impact on the traffic congestion are selected, like '逛街' (hang out), '出去吃饭' (eat out), '出发' (leave for) or words may indicate outdoor activities like '购物中心' (shopping center), '饭店'(restaurant) and '公园' (park).

2.2 Data Processing

The first task in this stage is to select the target dataset (analytic corpus). Tweets whose contents contain the words in the lexicon built in the previous stage are selected separately into four intermediate tables. Each of them contains tweets related to one subcategory: good weather, bad weather, outdoor activity, and traffic congestion. Next, a Ruby program computed the daily tweet counts for the four tables and saved them in CSV format.

2.3 Data Analysis

The core part of the whole work is data analysis stage. Since the mediation model is a causal model in nature, the causal relationships between weather conditions and traffic congestion, weather conditions and outdoor activity and outdoor activity and traffic congestion are determined first using Granger Causality Test. Afterwards, mediation analysis is conducted to model the relationships among weather conditions, outdoor activity and traffic congestion.

2.3.1 Granger Causality Test

The Granger Causality Test is a statistical hypothesis test for determining whether one time series is useful in forecasting another [11]. Basically, Granger Causality Test is used to determine if one time series variable X has causal relationships with another time series variable Y. In other words, it can be used to test the possibility to make prediction on one time series Y based on another time series X. Additionally, there is one basic assumption that must be fulfilled in order to use Granger Causality Test: time series X and time series Y should be stationary. Stationary time series keeps statistical properties such as mean, variance, autocorrelation, etc. constant over time. Therefore, meaningful sample statistics such as means, variances, and correlations with other variables can be obtained if the time series is stationary.

In this work, all four variables are time series variables since the values of each variable are collected at constant interval. Each data record is unique from other records and dependent only on time fields. Therefore, three groups of causal relationships are checked through this test: different weather conditions and traffic congestion, different weather conditions and outdoor activities as well as outdoor activities and traffic congestion.

As Fig. 2 shows, stationarity of the four time series variables are tested prior to the Granger Causality Test performed and only if the result shows independent variable and dependent variable are stationary, then the Granger Causality Test can be applied. Otherwise, the non-stationary variables must be transformed to stationary variables if either variable of the pair is non-stationary.

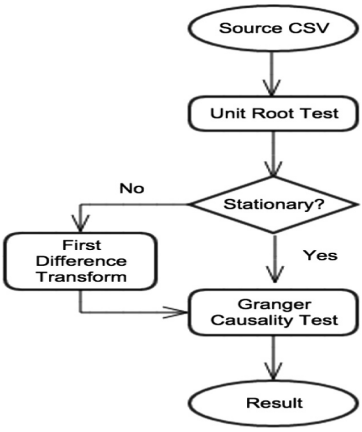

Fig. 2. Granger Causality Test mechanisms.

2.3.2 Mediation Analysis

As stated before, the mediation model is a causal model in nature. Therefore, the mediation analysis can be conducted once the causal relationships among weather conditions, outdoor activity and traffic congestion are determined. From the perspective of statistics, a mediation model is to identify an observed relationship between an

independent variable and a dependent variable via the inclusion of a third variable, known as mediator. Besides the direct relationship between the dependent variable and the independent variable, mediation model hypothesizes that the independent variable influences the mediator which in turn influences the dependent variable [15]. In other words, the mediator transmits the effect of an independent variable on a dependent variable. As demonstrated in Fig. 3, the effect of the independent variable on the dependent variable may be mediated by a mediator while the direct effect still remains.

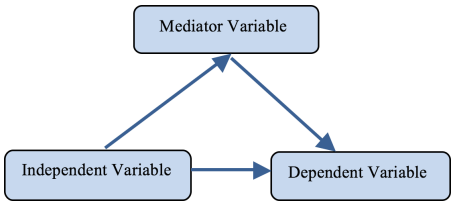

Fig. 3. A sample of mediation model

As against Fig. 3, the corresponding relationship in this work is: weather condition refers to the independent variable, traffic congestion refers to dependent variable and outdoor activity refers to the mediator.

To determine if outdoor activity is the mediator in this model, there are three major steps [16]: Step 1, Regress traffic congestion on the weather condition to ensure that weather condition is a significant predictor of traffic congestion. Step 2, Regress outdoor activity on the weather condition to confirm that weather condition is a significant predictor of outdoor activity. Step 3, Regress traffic congestion on both outdoor activity and weather condition to ensure that outdoor activity is a significant predictor of traffic congestion while controlling for the weather condition.

Additionally, the Sobel Test [17] is also implemented to determine whether the reduction in the effect of weather condition (the independent variable) on traffic congestion (the dependent variable), after including outdoor activity (the mediator) in the model, is a significant reduction and therefore whether the mediation effect is statistically significant.

3 Results and Discussion

The stationarity of these four variables are tested by Augmented Dickey–Fuller test (ADF) [12] and the results indicate that good weather and bad weather are stationary time series while outdoor activity as well as traffic congestion are non-stationary time series where the first difference of them are stationary. Consequently, the modified first differences of these non-stationary variables are used in the further analysis.

The Pairwise Granger Causality Tests are conducted using EViews, having the stationarity of both dependent and independent variables fulfilled. The test results demonstrate that prediction can be made on traffic congestion based on the good weather time series with the P value (probability) of the null hypothesis at 3.E−06.

Likewise, there are causal relationships between good weather and outdoor activity with P value of the null hypothesis at 8.E−09 as well as outdoor activity and traffic congestion with P value at 4.E−05. On the contrary, bad weather time series can be used to predict the traffic congestion (P value 0.0028) while it shows no obvious causal relationships with outdoor activity according to the test result (P value 0.0659).

Figure 4 illustrates the positive correlation between occurrence of daily tweets mentioning good weather and traffic congestion and outdoor activities. Each data point represents a day. On having social media can reflect people's opinions or thoughts to some extent, it can be inferred that there are more chances for people talking about traffic congestion when the weather is good and that people are more likely to do outdoor activities when the weather is good.

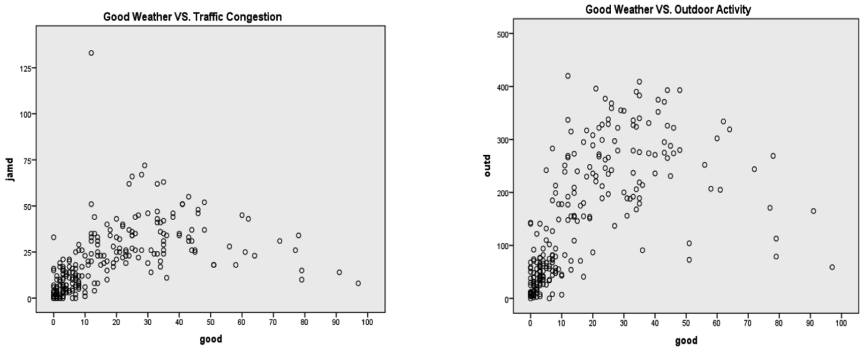

Fig. 4. Scatter chart of good weather vs. traffic congestion and outdoor activity

Similarly, the scatter chart in Fig. 5 demonstrates the positive trend between outdoor activity and traffic congestion. In accordance with the common sense, it can be seen clearly that more outdoor activities will cause more traffic congestions.

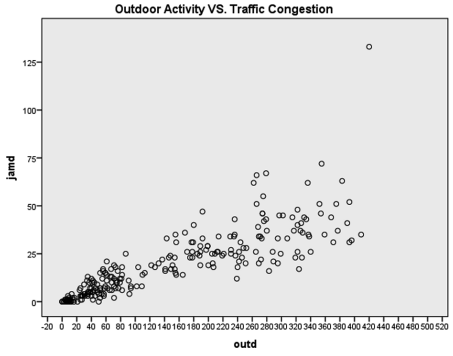

Fig. 5. Scatter chart of outdoor activity vs. traffic congestion

On having mediation model is a causal model in nature, the mediation model of good weather, outdoor activity and traffic congestion is then constructed after determining the causal relationships among these three time series variables. Based on the method to test whether outdoor activity can act as a mediator between good weather and traffic congestion, the regression results suggest that the significance of good weather on predicting traffic congestion decreases greatly when the outdoor activity intervenes. In addition, the Sobel Test result also indicates the mediation effect of outdoor activity is statistically significant with respect to the relationship between good weather and traffic congestion. Therefore, outdoor activity carries the effect of good weather on traffic congestion, which means it is a mediator in the mediation model shown in Fig. 6.

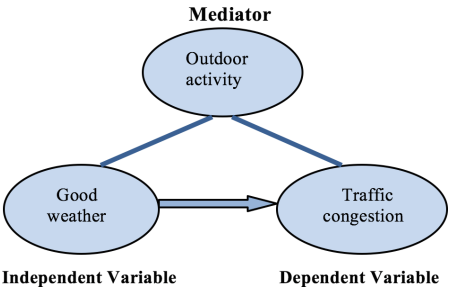

Fig. 6. Mediation model of good weather, traffic congestion and outdoor activity

4 Final Remarks

A practical approach that combines data mining techniques with statistical analysis is proposed to study the relationship between weather conditions and traffic congestion based on social networks. Following the general data mining procedures, statistical analysis can be incorporated with the final analysis stage. By crowd sensing the weather conditions and traffic congestion in Beijing using the proposed approach, it has been proved that good weather leads to traffic congestion and the direct cause is outdoor activity. This work provides a promising way to discover latent relationships between various activities in a smart city.

References

1. Derek, D., Karl, S., Swapna, S.G., Aldo, D.: Social media enabled human sensing for smart cities. AI Commun. **29**(1), 57–75 (2015)
2. Anjaria, M., Guddeti, R.M.R.: Influence factor based opinion mining of Twitter data using supervised learning. In: COMSNETS, January 2014, pp. 1–8 (2014)
3. Wu, X., Xie, F., Wu, G., Ding, W.: Personalized news filtering and summarization on the web. In: Proceedings of the 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence, Washington, DC, USA, pp. 414–421 (2011)

4. Nathan, K.C., Amanda, L.G.: Health behavior interventions in the age of Facebook. Am. J. Prev. Med. **43**(5), 571–572 (2012)
5. Fisher, J., Clayton, M.: Who gives a Tweet: assessing patients' interest in the use of social media for health care. Worldviews Evid Based Nurs. **9**(2), 100–108 (2012)
6. Wang, S.L., Paul, M.J., Dredze, M.: Social media as a sensor of air quality and public response in China. J. Med. Internet Res. **17**(3) (2015)
7. Pan, B., Zheng, Y., Wilkie, D., Shahabi, C.: Crowd sensing of traffic anomalies based on human mobility and social media. In: ACM SIGSPATIAL GIS pp. 344–353 (2013)
8. Cools, M., Moons, E., Wets, G.: Assessing the impact of weather on traffic intensity. Weather Clim. Soc. **2**, 60–68 (2010)
9. Zeng, K., Liu, W.L., Wang, X., Chen, S.H.: Traffic congestion and social media in China. Intell. Syst. **28**(1), 72–77 (2013)
10. Chifor, B.C., Bica, I., Patriciu, V.V.: Sensing service architecture for smart cities using social network platforms. Soft Comput. **21**, 1–10 (2016)
11. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica **37**(3), 424–438 (1969)
12. Dickey, D.A., Fuller, W.A.: Distribution of the estimators for autoregressive time series with a unit root. J. Am. Stat. Assoc. **14**(366), 427–431 (1979)
13. Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time Series Analysis: Forecasting and Control, 3rd edn. Prentice Hall, Englewood Cliffs (1994)
14. Anderson, O.: Time Series Analysis and Forecasting: The Box-Jenkins Approach. Butterworths, London (1976)
15. MacKinnon, D.P.: Introduction to Statistical Mediation Analysis. Erlbaum, New York (2008)
16. MacKinnon, D.P., Fairchild, A.J., Fritz, M.S.: Mediation analysis. Annu. Rev. Psychol. **58**, 593–614 (2007)
17. Sobel, M.E.: Asymptotic confidence intervals for indirect effects in structural equation models. Sociol. Methodol. **13**, 290–312 (1982)