2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Mining Twitter and Taxi Data for Predicting Taxi Pickup Hotspots

Sankarshan Mridha*, Sayan Ghosh†, Robin Singh‡, Sourangshu Bhattacharya§ and Niloy Ganguly‖
Indian Institute of Technology, Kharagpur - 721302, WB, India
{*sankarshan,†sgdgp,‡robinsingh}@iitkgp.ac.in, {§sourangshu,‖niloy}@cse.iitkgp.ernet.in

***Abstract*—In recent times, people regularly discuss about poor travel experience due to various road closure incidents in the social networking sites. One of the fallouts of these road blocking incidents is the dynamic shift in regular taxi pickup locations. Although traffic monitoring from social media content has lately gained widespread interest, however, none of the recent works has tried to understand this relocation of taxi pickup hotspots during any road closure activity. In this work, we have tried to predict the taxi pickup hotspots, during various road closure incidents, using their past taxi pickup trend. We have proposed a two-step methodology. First, we identify and extract road closure information from social network posts. Second, leveraging the inferred knowledge, prediction of taxi pickup hotspot is done near the activity location with an average accuracy of $\sim 86.04\%$, where the predicted locations are within an average radius of only 0.011 mile from the original hotspots.**

***Keywords*—Social Network, Transportation Network, Classification, Taxi Pickup Hotspot Prediction, Hotspot Relocation**

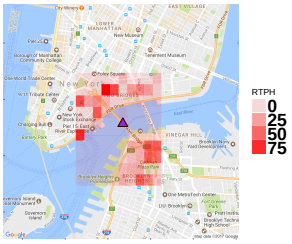
I. Introduction

In large cities taxi services normally maintain hotspots - the knowledge of which helps daily/experienced commuters to quickly and easily avail taxi service. Such hotspots may get disturbed during a road blockage (e.g. closure due to maintenance) around that area, whereby taxis dynamically build new hotspots. These new hotspots may be far away from older hotspots, hence causing inconvenience to the commuters as well as resulting in loss of business and time of taxi drivers. A service which can predict the new hotspots and accordingly suggest its users to the most convenient new hotspot would be of tremendous value to both experienced and new commuters.

For such a service, we need to automatically identify road closure events as well as gather information regarding traffic patterns and predict taxi hotspot location. The road closure events can be detected from social media posts, and information regarding taxi traffic patterns are released by authorities in various cities, e.g. Porto, Boston, Chicago, New York, etc., and also by taxi hiring companies (e.g. Uber[1]). In this paper, we concentrate our study on New York City (NYC), for these informations are freely available. The NYC Department of Transportation (twitter handle: NYC_DOT) regularly post tweets about various events in New York, including road-closure (e.g. Figure 1-(a)). The New York City Taxi and Limousine Commission routinely releases information regarding trips of yellow taxi, their pick-up and drop-off points from which one can calculate the hotspots in the city. Figures 1-(b) and (c) show the Taxi Pickup Hotspots (RTPH), before and after a road blocking event. A close look at the two pictures shows the difference in heat-map and indicates the change in taxi pickup hotspot.

(a)

(b)



(c)

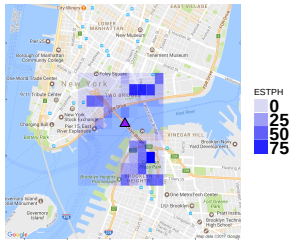


Fig. 1: (a) Tweet notifying lane closure activity in Brooklyn Bridge. Heatmap showing (b) Regular Taxi Pickup Hotspot (RTPH) before and (c) Event Specific Taxi Pickup Hotspot (ESTPH) after road closure activity around Brooklyn Bridge (purple triangle).

Summing up, we present a two step methodology to tackle this problem of finding taxi pickup hotspots by analysing posts from Online Social Network (OSN). We have used two datasets: (1) the tweets posted by New York City Department of Transportation (twitter handle: NYC_DOT[2]) for the year 2015 for inferring the social cue and (2) the New York City Taxi cab data[3] for the year 2015. Our two step process includes two broad tasks: (a) Natural Language Processing (NLP) task for information extraction and (b) Map-based task for hotspot prediction during road closure activity using the extracted information. In the first step, we present method for NLP tasks, classify eventful tweets, followed by extraction of meaningful information on imminent traffic disturbances by using the notification tweet posted by NYC_DOT. After comparing the road closure incidents between, the year 2015 and 2016, we observe that among all the road closure locations,

[1]https://movement.uber.com/cities

ASONAM '17, July 31 - August 03, 2017, Sydney, Australia

ACM ISBN 978-1-4503-4993-2/17/07?/$15.00
http://dx.doi.org/10.1145/3110025.3110106

[2]https://twitter.com/NYC_DOT
[3]http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

$\sim 40.47\%$ (section II-B) event locations are repeated across years. Second we go for Map-based task where using the extracted information, we try to predict the taxi pickup hotspot around the activity location using the NYC Taxi cab data.

To summarize the key points of our works as follows: (i) We show the method to learn the taxi pickup hotspots combining the online and offline data source (section IV,VI) using a two step process. (ii) Our approach is successfully classifying the relevant and non-relevant traffic tweet with an average accuracy of $\sim$**94.55%**, precision $\sim$**0.94** and recall $\sim$**0.94** (section V). (iii) Following the classification, we have successfully extracted date-time and location information from $\sim 80\%$ of relevant tweet (section V). (iv) In section VII we show how fair we are doing in predicting the pickup hotspot. It shows that our average relative absolute error (RAE) is only $\sim$**13.96%** and root relative squared error is $\sim$**22.33%** for locations with higher number of road closure incidents.

II. Importance and Feasibility of the Problem

A. Importance of the Problem

Death of Old Hotspot: Here, some of the old hotspots no longer remain as hotspots, as the passenger pickup count drops significantly in those places. In figure 1-(c), some of the old hotspots become obsolete, due to road closure activity at Brooklyn Bridge, compared to figure 1-(b). This trend is also occurring in other places of New York City.

Birth of New Hotspot: As a side effect of the road closure incident, new hotspots also appear in the locality, where previously no passenger pickup was done – birth of new hotspot. In figure 1-(c), some new event specific taxi pickup hotspots (ESTPH) have come up around Brooklyn Bridge, where previously no hotspots were there (figure 1-(b)).

Hotspot Relocation: We also found that the minimum shift (in mile) of an old hotspot to a new hotspot for each of these locations due to such road blocking activity. The minimum shift is reported as ~ 0.104 mile. This shows, at the arrival of road closure events, at least how far the hotspots move from its usual places.

B. Feasibility of the Solution

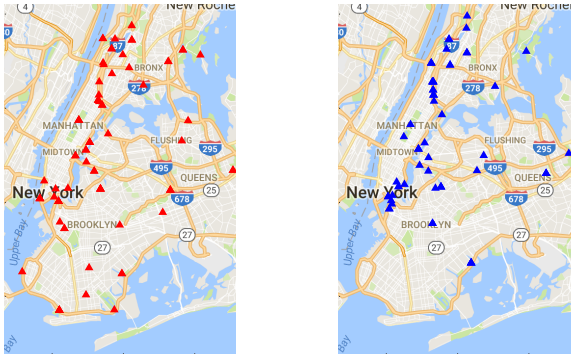
Total incidents: After analysing the data it's found that there are 367 and 417 road closure incidents in New York City for year 2015 and 2016 respectively. These incidents are spanned over 67 locations in 2015 and 42 locations in 2016.

Affected region and overlap: To understand their geographical spreading, the coordinates of the incident locations are plotted on the global map. The figure 2 interprets that majority of the incident locations are densely spread over Manhattan. But rest are sparsely located over Bronx, Brooklyn and Queens. The analysis also finds that there is 40.47% repetition in incident locations between these two years. This is even clearly visible in the figure 2. For Manhattan, the geographic regions for the road closure incidents are very similar in 2015 and 2016. However for Bronx, Queens and Brooklyn, we see that the incidents occur in few new locations. This tells that, these repeated road activities are mainly affecting Manhattan.

III. Related Work and Dataset

Computing with Homogeneous OSN data: Event extraction from social media data is well studied [1],[2],[3]. There are also works on finding out fruitful traffic insights from the social media [4],[5]. Pathak et al. [4] shows how data from social network like facebook can deliver useful urban traffic information using multi-class classification technique to categorize traffic incidents. Tejaswin et al. [5] uses the spatio-temporal tweet feed data for traffic incidents clustering and prediction using Random Forest model.

Computing with Heterogeneous Urban data: Wang et al. [6] and Lécué et al. [7] extend the traditional traffic sensing paradigm by coupling various sensors (e.g. GPS) with OSN, whereas Daly et al. [8] and Rashidi et al. [9] use heterogeneous mixture of social media data to infer different traffic attributes. Pan et al. [10] explores the traffic anomalies in the transportation data. Finally Wu et al. [11] combines ubiquitous data sources for explaining daily traffic trend by finding traffic correlation with these external data sources.



(a) (b)

Fig. 2: Distribution of road closure event locations in New York city for year (a) 2015 (b) 2016. The major locations are distributed mainly in Manhattan with good yearly overlap.

IV. Tweet Acquisition and Preprocessing

A. Tweet data-source and classification

Twitter data-source: We have crawled the twitter timeline of New York City Department of Transportation for year 2015 and 2016 using twitter REST API[4]. The 2015 dataset contains tweet feed from 25^{th} Dec 2014 to 12^{th} Dec 2015. For 2016, the dataset is from 4^{th} Jan 2016 to 12^{th} Dec 2016. The total number of tweets for year 2015 and 2016 are 2981 and 2966 respectively.

Tweet classsification: We use standard techniques for classification of relevant tweets. We assume that $tf_i = \{f_{i1}, f_{i2}, \ldots, f_{iN}\}$ is the tweet-feature vector for i^{th} tweet, where f_{ik} denotes the k^{th} feature value for i^{th} tweet. We describe the features used in section V. Let $C_i \in \{0, 1\}$ denote the class label capturing whether i^{th} tweet is relevant or not. Here by relevant, we mean one containing information about road closures. Our final objective is, to classify a given unseen tweet as relevant or not. In summary our input, outputs are as follows: *Input:* A fixed set of tuples having tweet-feature vector and corresponding class labels $[(tf_1, C_1), (tf_2, C_2), \ldots, (tf_M, C_M)]$ and *Output:* A binary classification model $\boldsymbol{CR} : tf \rightarrow C$ where $\boldsymbol{CR}$ classifies the tweet as relevant or not.

[4]https://dev.twitter.com/rest/public

	2015			2016		
	Acc(%)	**P**	**R**	**Acc(%)**	**P**	**R**
SMO	92.82	0.925	0.928	94.97	0.951	0.950
RF	93.85	0.936	0.939	94.94	0.948	0.949
BG	**93.86**	**0.938**	**0.939**	**95.24**	**0.951**	**0.952**
NB	93.09	0.931	0.931	94.97	0.951	0.950

TABLE I: Tweet classification performance: Accuracy (Acc), Precision (P), Recall (R) for 2015 and 2016 tweets.

B. Inferring Road Closure Information

Extracting Closure Location: First, we do parts of speech (POS) tagging for the tweet and select only the words having NNP (proper and singular noun) tag – first set of candidate words. Second, we filter out all the hashtags which are used for mentioning the name of the event locations – second set of candidate words. Combining these two sets we build the list of potential location names. Next, we query the location names (using python *geopy* library) to find the coordinates which these words might represent and choose only those which are located inside NYC.

Extracting Time: We extract all the text from the relevant tweet which are succeeded by "am" or "pm" using the regular expression: "$*(am|pm)*$". We use regular expression based rules to extract *date*, *time* information from other numeric and alphanumeric strings.

Extracting Date: To identify the *date* information we check for the presence of *absolute mention* and *relative mention* (e.g. *today*, *tomorrow*) of the date. We use "$tomorrow|tmrw|yesterday|tonight|today$" as the regular expression for this task. If a relative reference for *date* is found then the date for the event is calculated using the posting time of the tweet. We also take the help of *conjunction words*. These words help to identify the duration of the event and draws relation between different dates and time. Types of conjunction words used in our regular expression are "$through|thru|to|between|btn|\&|and|-$".

V. Results: Tweet Mining

Ground Truth Preparation: The original tweets were not labelled as relevant or non-relevant. So 2 person manually annotate them as relevant and non-relevant. The kappa value for inter annotator agreement between them is 0.76. Next we move on to feature selection task for tweet classification.

Feature Selection for Classification: The relevant tweets contain four key patterns, which helps to correctly classify them. These key features are as follows: (i) *Key Words:* For all the tweets, word frequency distribution analysis is done for both relevant and non-relevant tweets separately, after normalizing them (removing stop words followed by applying stemming). Finally we select top-20 key words as one of our feature. (ii) *Date:* As it's a notification regarding an upcoming incident, they always mention about the *date* of the event, which makes the tweet distinct from the non-informative one. (iii) *Time:* These road activities are also specific to a particular *time* say from $10:00$ am to $3:00$ pm. This adds more value to the tweet and helps us in correctly identifying them. (iv) *Name of the Day:* Finally along with *date* and *time*, the *day* of the week has also been mentioned separately which is also added in the feature list.

Accuracy of classifiers: Tables I shows the classification accuracy, using $10-$fold cross validation, for tweets of year 2015 and 2016, respectively. Since, we are trying to retrieve all the relevant tweets, therefore our main objective is to build a high recall system. We see for both 2015 and 2016 tweets, Bagging gives the highest average precision $\sim$**0.945**, recall $\sim$**0.945** and accuracy values $\sim$**94.55%**.

Information Extraction Result: We are able to extract the *date* and *time* information from 84.7% and 88% relevant tweets for year 2015 and 2016 respectively. However in case of location extraction the result is around 73.23% and 72.1% for year 2015 and 2016 respectively. The poor performance is due to image based notifications, which we are not processing, or name ambiguity – different shorter name of the same location.

VI. Predicting Taxi Pickup Hotspots

Problem Formulation: Assume that at location l, N road closure events $\{e_1, e_2, \ldots e_N\}$ have occurred at times $\{t_1, t_2, \ldots, t_N\}$ respectively. Given this information we want to predict the taxi pickup *hotspot* around the event location l for a new event. To solve this problem, first we consider a grid G of size $k \times k$ in the neighbourhood centring the location. For each grid cell $g \in G$, we select list of features F and create a feature vector $v_g = \{f_{g1}, f_{g2}, \ldots, f_{gk}\}$ using transportation data, where f_{gi} denotes the i^{th} feature value for grid cell g. We have the pickup count y_g for grid cell g that denotes the total number of pickups that occurred when the road-closure incident happened at location l at time t. Next combining all the feature vector we build our feature matrix $\boldsymbol{A} = [v_1, v_2, \ldots, v_N]^T$ and target vector $\boldsymbol{y} = [y_1, y_2, \ldots, y_N]^T$. Finally, we apply different supervised learning technique to estimate $\boldsymbol{y}$.

Feature Engineering: We use the following features: (1) *LCD:* A grid structure around the incident location has been constructed and the *location coordinates* (LCD: lat, lon) of these event areas are included in the feature list. (2) *PHN:* We include the *pickup history of neighbourhood* (PHN) in our feature set as the total pickup counts for the past W days that occurred during the same time period at location l. (3) *TOD:* Following the hourly trip distribution pattern in [12], we include *time of the day* (TOD) in our feature list. (4) *DOW:* Next we focus to find the effect of different *day of the week* (DOW) over the daily trip count. (5) *DFL:* Finally we assume that the effect of traffic disturbance gradually fades away from the epicentre – inclusion of *distance from location* (DFL) in the feature list.

VII. Results: Taxi Pickup Hotspots

Input Taxi Data: For the transportation data, the New York City (NYC) Taxi data for the year 2015 has been used. This dataset is made public by NYC Taxi and Limousine Commission. They provide the data for all yellow taxi trips made in NYC since 2009. However we have used this data for the year 2015 only. After initial data processing, cleaning and filtering it has ~ 163 million taxi trip data for year 2015.

Local Grid: Once road closure locations have been identified, we go for a grid based approach for learning taxi pickup hotspot following the idea of Tejaswin et. al. [5]. We use the idea of localized grid around the road closure locations and estimate the pickup count in these grids during the event. For each location we take a square region of size 2×2 sq. miles centring the location. Next we break the region in size of

Location	SVR			LR			RF			BG		
	Cor	**RAE**	**RRSE**	**Cor**	**RAE**	**RRSE**	**Cor**	**RAE**	**RRSE**	**Cor**	**RAE**	**RRSE**
Battery Park	0.98	13.90	22.05	0.98	14.95	22.02	**0.98**	*11.30*	**18.69**	0.98	12.11	19.76
Manhattan Br	0.97	16.87	26.03	0.97	14.41	24.28	**0.97**	*15.31*	**22.67**	0.965	16.82	26.32
Brooklyn Br	0.93	23.03	37.89	0.93	24.26	37.44	**0.97**	*17.06*	**24.83**	0.96	17.97	27.23
Roosevelt BR	0.97	13.90	25.24	0.97	14.41	25.81	**0.97**	*12.18*	**23.13**	0.97	19.12	25.58
QueensBoro Br	**0.98**	**14.05**	**21.84**	0.97	14.31	24.28	0.96	17.46	28.78	0.95	19.40	33.46
WillsBurg Br	**0.93**	**28.31**	**36.78**	0.93	28.62	37.63	0.87	33.68	48.67	0.87	34.03	49.88
WillAve Br	**0.74**	**35.53**	61.80	0.79	44.32	61.33	0.77	42.63	64.66	0.79	41.45	61.44

TABLE II: Result showing hotspot learning by different prediction model– Support Vector Regression (SVR), Logistic Regression (LR), Random Forest (RF), Bagging (BG) where Correlation Value (Cor) ranges between (0,1) and relative absolute error (RAE) and root relative squared error (RRSE) are in percentage (%).

0.2×0.2 sq. miles. Now we move on to the hotspot learning task as discussed in section VI. For each grid location we build the feature matrix $\boldsymbol{A}$ and the target vector $\boldsymbol{y}$ and learn our model. After, estimating the pickup count for each cell, we apply minimum pickup count threshold as 3 for labelling each cell as hotspot or not.

Predicting Taxi Pickup Hotspot: After we select the features, we build our feature vector $v_g = \{LCD_g, PHN_g, TOD_g, DOW_g, DFL_g\}$ for each neighbouring region. Next we build our feature matrix $\boldsymbol{A}$ and target vector $\boldsymbol{y}$ as discussed in section VI and apply different learning algorithm for predicting the pickup count in the neighbouring locations. To measure the accuracy of our learning system, we calculate the Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE). If for i^{th} observation, the predicted value is $\hat{y}_i$ and the observed value is y_i then: $RAE = \frac{\sum_{i=i}^{N} |\hat{y}_i - y_i|}{\sum_{i=i}^{N} |\bar{y}_i - y_i|}$ and $RRSE = \sqrt{\frac{\sum_{i=i}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=i}^{N} (\bar{y}_i - y_i)^2}}$. Finally RAE and RRSE are multiplied with 100 to set the scale in $0 - 100$. Here, the values are normalized by how much y differs from it's mean value.

Table II shows the performance of the learning technique on these locations using $10-$fold cross validation. We see that Random Forest (RF) and SVR are giving the most promising result. The table shows that Random Forest (RF) gives good result for Battery Park, Manhattan Bridge, Brooklyn Bridge and Roosevelt Bridge compared to the other three model. Here the average correlation score between the observed and the predicted value is 0.975. Whereas the average RAE is $\sim 13.96\%$, RRSE is $\sim 22.33\%$ for these four locations. For QueensBoro Bridge, SVR is giving comparatively better result compared to Random Forest. Here the average correlation between the observed and the predicted is 0.98, RAE is 14% and RRSE is 21.84%. But for Williamsburg Bridge and Wills Avenue Bridge the result is not that good where average correlation is only 0.835. The reason for such weak performance of the model is due to the smaller data size as comparatively fewer number of road closures incidents occurs in these places.

VIII. Conclusion

In this paper, we study the problem of taxi pickup hotspot relocation due to road closure events, by analysing the event notification in the online social network and taxi transportation data. We show that organic process of relocation of hotspots is predictable. We design a novel two-step process for automatically detecting and locating road closures from traffic notifications posted on twitter, and then predicting locations of new hotspots. This can used to build a hotspot recommendation service, which will be helpful to both newcomers and residents of the city.

References

[1] E. Benson, A. Haghighi, and R. Barzilay, "Event discovery in social media feeds," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 389–398.

[2] A. Ritter, O. Etzioni, S. Clark *et al.*, "Open domain event extraction from twitter," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1104–1112.

[3] H. Becker, D. Iter, M. Naaman, and L. Gravano, "Identifying content for planned events across social media sites," in *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 2012, pp. 533–542.

[4] A. Pathak, B. K. Patra, A. Chakraborty, and A. Agarwal, "A city traffic dashboard using social network data," in *Proceedings of the 2Nd IKDD Conference on Data Sciences*, ser. CODS-IKDD '15. New York, NY, USA: ACM, 2015, pp. 8:1–8:4. [Online]. Available: http://doi.acm.org/10.1145/2778865.2778873

[5] P. Tejaswin, R. Kumar, and S. Gupta, "Tweeting traffic: Analyzing twitter for generating real-time city traffic insights and predictions," in *Proceedings of the 2nd IKDD Conference on Data Sciences*. ACM, 2015, p. 9.

[6] Y. Wang, "Socializing multimodal sensors for information fusion," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 653–656.

[7] F. Lécué, R. Tucker, V. Bicer, P. Tommasi, S. Tallevi-Diotallevi, and M. Sbodio, "Predicting severity of road traffic congestion using semantic web technologies," in *European Semantic Web Conference*. Springer, 2014, pp. 611–627.

[8] E. M. Daly, F. Lecue, and V. Bicer, "Westland row why so slow?: fusing social media and linked data sources for understanding real-time traffic conditions," in *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 2013, pp. 203–212.

[9] T. H. Rashidi, A. Abbasi, M. Maghrebi, S. Hasan, and T. S. Waller, "Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges," *Transportation Research Part C: Emerging Technologies*, vol. 75, pp. 197–211, 2017.

[10] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," in *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2013, pp. 344–353.

[11] F. Wu, H. Wang, and Z. Li, "Interpreting traffic dynamics using ubiquitous urban data," in *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2016, p. 69.

[12] X. Zhan, S. Hasan, S. V. Ukkusuri, and C. Kamga, "Urban link travel time estimation using large-scale taxi data with partial information," *Transportation Research Part C: Emerging Technologies*, vol. 33, pp. 37–49, 2013.