STAR: Real-time Spatio-Temporal Analysis and Prediction of Traffic Insights using Social Media

Deepali Semwal, Sonal Patil, Sainyam Galhotra, Akhil Arora, and Narayanan Unny
Xerox Research Centre India, Bangalore
{deepali.semwal, sonal.patil, sainyam.galhotra, akhil.arora, narayanan.unny}@xerox.com

ABSTRACT

The steady growth of data from social networks has resulted in wide-spread research in a host of application areas including *transportation*, health-care, customer-care and many more. Owing to the ubiquity and popularity of transportation (more recently) the growth in the number of problems reported by the masses has no bounds. With the advent of social media, reporting problems has become easier than before. In this paper, we address the problem of efficient management of transportation related woes by leveraging the information provided by social media sources such as – Facebook, Twitter etc. We develop techniques for viral event detection, identify frequently co-occurring problem patterns and their root-causes and mine suggestions to solve the identified problems. We predict the occurrence of different problems, (with an accuracy of $\approx 80\%$) at different locations and times leveraging the analysis done above along with weather information and news reports. In addition, we design a feature-packed visualization that significantly enhances the ability to analyse data in real-time.

1. INTRODUCTION

Transportation related services are one of the major contributors in ensuring sustainable economic growth of a country. However, emerging markets like India suffer from the many repercussions of rapid development – infrastructural insufficiency in traffic management is one such issue. The road network alone forms 40% of the total available transportation modes in India. These percentages are similar in other countries as well. Given this huge percentage, large number of problems pertaining to accidents, traffic etc. have been a major concern. The revolution in the automobile industry with supports from liberalized economy has led to tremendous increase in the vehicle ownership levels which in turn has caused problems pertaining to accidents, traffic etc.. This has resulted in continuously evolving traffic characteristics on road networks, thus significant improvements are desired over the existing infrastructure.

CODS-IKDD'15, March 20-20, 2015, Bangalore, India.
ACM 978-1-4503-3616-1/15/03.
http://dx.doi.org/10.1145/2778865.2778872.

In this paper, we propose an approach which aims at providing a low cost solution for traffic management using the data available on social media and other sources like weather and news reports. We have devised novel techniques to monitor various issues in different parts of the metropolitan cities of the nation and flag them according to their severity. We analyse the temporal aspects of incoming posts to detect prevalent places and problems in a city at a point in time. We employ frequent pattern mining to uncover frequently co-occurring problems. Further analysis of the posts are done to report causes of the problems along with the popular suggestions from the citizens. The posts by users are used to suggest useful changes to the authorities, if any, in order to resolve the registered complaints. Finally, we use data from Facebook posts combined with multiple other sources to predict the next days problem in a given region with an high accuracy.

To complete this, we have devised a web based application as shown in Figure 1 to facilitate effective visualization of the interesting results mined by our techniques. This application monitors various traffic related issues reported by people over social media and analyses them to provide useful insights to the concerned authorities. The end goal of this work is to assist the traffic authorities of various cities in India, thus helping them address the urgent traffic issues faced by the masses.

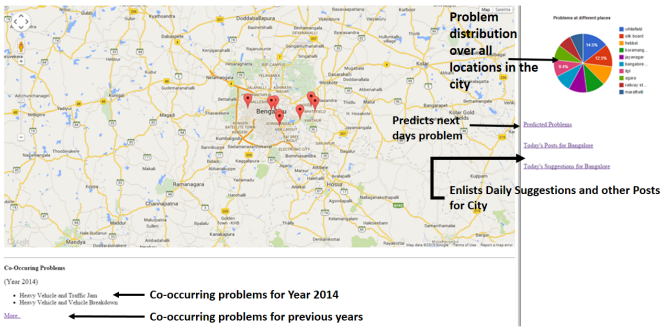

Figure 1: Snapshot of the system with different modules.

Our main contribution in this work would be the following. In our end to end system, we have developed some complex machine learning techniques for event/problem detection, sentiment analysis, location extraction and suggestion clas-

sification. We come up with a holistic tool to analyse the problems, come up with root-causes of these problems and recommend popular suggestions for the same. Finally, we introduce the temporal aspect of the data in our prediction model along with weather information and news data, and develop a novel next problem prediction algorithm.

2. RELATED WORK

Horvitz et al. [1] have devised techniques to analyse and predict traffic by deploying sensors, which though being interesting possess the disadvantage of being expensive. Ni et al. [2] propose a technique to combine social media data with traffic data available from sensors to predict traffic flow on special events. Yu et al. [4] have proposed different techniques leveraging the data from social media. Their prediction technique can be easily used in the field of marketing, elections, etc. but they have not discussed a way to detect the events which they want to track. Moreover in the traffic scenario, predicting using the twitter data alone is not justified as there are many extraneous factors affecting the flow of traffic e.g. rainfall, special events and many more.

Google API[1] analyses traffic by monitoring the network information available but have no provision to predict the amount of traffic in the near future. Traffline[2] is a web application which analyses the GPS on taxis, posts on twitter. Their event detection is only on the basis of inputs provided by volunteers. As mentioned on Wikipedia[3], they do not have any provision to predict the most severe problem. In this paper, we aim to bridge these gaps by detecting the occurrence of any event from Facebook data along with weather information to predict major problems.

3. SYSTEM DESCRIPTION

The system is based on different processing modules as depicted in Figure 3. We have used Facebook public page posts to get updates about the traffic. Along with Facebook data, other data sources such as weather data and news data are also contributing for the problem of prediction. The collected data is preprocessed to get the useful information and is stored in a MYSQL data store. The processed information is in turn used by different modules in the system. This end to end system is being devised for getting insights of traffic. It is evident that the system solves four main sub problems which are discussed below.

3.1 Data and Pre-processing

Data for the system is constituted of the social media feeds pertaining to the traffic of different cities. For each data input, we remove posts with content in languages other than English. We remove duplicate posts. We segregate these posts on coarse city basis and further analysis is done on the city level data. Furthermore, for each city, the data is partitioned on the basis of fine grained locations. A dictionary of city's fine grained location is populated by crawling the pages from official sites of the city.

3.2 Virality Detection

This module uses the problem information associated with the location for all days extracted from the social media

[1]https://developers.google.com/maps/
[2]http://www.traffline.com/
[3]http://en.wikipedia.org/wiki/Traffline

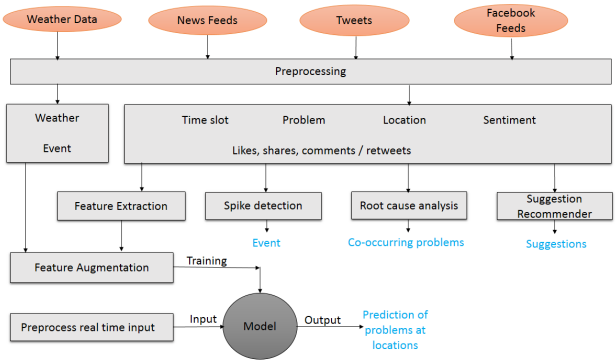

Figure 3: Overview of our System.

data. The basic intuition behind this module is that if there is some event happening at some location, people are more likely to discuss about it. Any problem at a location that is persisting over time can lead to severe traffic issues. To identify an event occurring over a period of time, we detect a spike in the number of posts at particular location over that time period.

A number of posts for a specific problem at some location, L can be modelled as series n_0, n_1, ..., n_W where n_0 is the score of the problem at location L to start with, n_W is the score of posts at the W^{th} day from start. The score is the count of posts mentioning the problem at particular location. This score of problems is computed by considering the severity of the posts, their impact which is measured by its likes, comments, shares, etc. W is the window size which will be moving over time by t days. The variables W and t can be varied as per the requirement and availability of resources. The total number of posts over 0^{th} window of size W say T_0 and 1^{st} window is T_1. If there is a difference between T_1 and T_0 which is greater than threshold α then the spike is detected.

$$T_1 = n_0 + n_1 + ... + n_W \quad (1st\ time\ window)$$
$$T_2 = n_0 + n_1 + ... + n_W \quad (2nd\ time\ window)$$
$$if(T_2 - T_1 > \alpha)$$
spike detected

Once the spike is detected the value of T_1 is noted and the window is moved ahead. T_1 represents the beginning of the spike. The spike will persist even if the consecutive window (T_i and T_{i+1}) has less difference than the threshold α, but T_{i+1} is consistently greater than T_1. So, till the time, T_{i+1} is in rising pattern as compared to T_1, the spike exists. Once the (T_{i+1}) goes below T_1, it means the spike has actually started decreasing. But it may happen that it can again pick up. So assuming d days as supplementary as a part of spike duration, we compute the window differences of the current and T_0 window (window before the spike begins). If after d days the difference is still less than α then the spike is considered to have finished. If the difference increases again than α then the spike is continued.

The posts associated with the detected spike are then analysed to evaluate their sentiments. Presence of a negative sentiment, indicates the presence of some event/problem which has an impact on traffic of that location. So the problem will be detected and based on the duration of the spike and impact score of the posts during that duration,

	Data					Sentiment Analysis	Real Time	Prediction
	Traffic Sensor data	Event Information	Weather Data	Social Media Data	GPS / Mobile Data			
Horvitz et al.	✓	✗	✗	✗	✗	✗	✗	Traffic
Ni et al.	✓	✓	✗	✓	✗	✗	✗	Traffic flow on special events.
Yu et al.	✗	✗	✗	✓	✗	✓	✗	Event based tracking
Google API	✗	✗	✗	✗	✓	✗	✓	✗
Traffline	✗	✗	✗	✓	✓	✗	✓	✗
STAR*	✗	✓	✓	✓	✗	✓	✓	**Major problems pertaining to traffic**

Figure 2: Comparison between various existing and proposed solution.

the severity of the problem will be calculated. The impact score of a post is defined as a function of the reliability of the user posting it, number of likes, shares, comments and their sentiments.

Also, if we find spike at some other location during the same/overlapping duration, we will check the latitude and longitude of these locations to check if they are in vicinity. If they are found to be closely connected then we can find the seed location of the problem by analysing the posts. The influential seed can be identified as the origin of the problem.

3.3 Co-occurring problems and causes

A majority of people mentioning problems (e.g. traffic jam, no parking) in their tweets, posts are likely to mention the causes (e.g. Heavy vehicle, break down or u-turn) for the problem. These causes might remain same or vary with the change in location. Further, a post may contain mentions of more than one problem occurring at a location. Here, we use Apriori algorithm [3] to identify these frequently co-occurring problems along with identifying the causes for the problems from these posts.

We maintain two sets consisting of pre-identified problems $\{p_1, p_2, p_3, ...\}$ and causes $\{c_1, c_2, c_3, ...\}$. We are now interested in identifying the commonly co-occurring items across and within these sets. We create a record of co-occurring problems and causes corresponding to each post $\{p_1, c_1\}$, $\{p_2, p_3\}$, ... etc. This data constitutes the input for Apriori Algorithm for finding the frequent patterns. We identified these co-occurring problems/causes for a data of over four years (See Table 1).

3.4 Suggestions Recommender

The events detected by *virality detection* module is provided to *co-occurring problems identification* module to identify the event that has resulted into sudden spike over time period. The corresponding posts over the time period of the spike are then analysed to get suggestions out. This module also helps in providing daily suggestions provided by the people for traffic personnel. A simple rule based classifier is used to classify the posts into suggestions and other posts.

3.5 Problem Predictor

In this module, we predict the occurrence of major problems at a given instant of time. It can be seen that the occurrence of a problem in near future can be attributed to following four reasons,

- A particular location faces the problem every week on the same day around the same time (eg. traffic jam during working hours i.e. morning time).
- The mishappenings/event which happened few days back trigger the issue.
- The occurrence of some public event leads to traffic jam and in turn problems faced by that location.
- The unusual change in weather of a particular location can also affect the normal course of traffic.

To predict the prevalence of a problem at a particular location, we use these information to devise features for training a classifier. Suppose we want to predict the problem faced by a location l on day d. The first feature set extracted corresponds to the scores of each problem for the location l on days $d-7$, $d-14$ and $d-21$. The score of each problem is calculated the same way, we calculate in the first module. The second feature set extracted corresponds to the scores for days $d-1$, $d-2$, ..., $d-6$ on the same lines. The third feature set extracted represents the data for whole one month history. All these features' weight vary as per relevance. Since we are trying to predict current trend, latest information will have more weight. Hence, scores of last week gets higher weights than that of whole month. The fourth feature tries to incorporate the occurrence of any public event around any place like rally, road show, etc. To capture such event we use the data available on news websites. The occurrence of an event can be judged by mining the number of people going to the event from the Facebook page of the event. If this number is more than a threshold, then it would affect the course of traffic for that location. The fifth feature tries to capture the effect of weather to predict the traffic patterns as many roads get clogged on the days of heavy rainfall leading to some specific problems. We use random forest classifier to predict most dominant issue for the next day. SMOTE is applied on the data set to reduce class imbalance which improved the accuracy further (Refer section 4).

4. EXPERIMENTAL EVALUATION

Evaluation is conducted on two levels system validation and technique performance. System validation includes validating the event detected by the module against an actual event. Technique performance includes the accuracy check of the classifier used to get its appropriateness for the domain.

The results shown here, are for different locations of Bangalore city. Figure 4 displays the number of posts for the area 'Whitefield, Bangalore' during the period '20-11-2014' and '04-12-2014'. It can be seen that there is a spike in the number of posts around Nov, 27. We have identified from the discussion on posts that people were talking about 'Whitefield rising' event which was change in route, implemented from '24-11-2014'. We have analysed the sentiments of people about this particular event and identified that about 60% people were talking about this were positive. So, this change has been taken positively and improved the traffic management of Whitefield area.

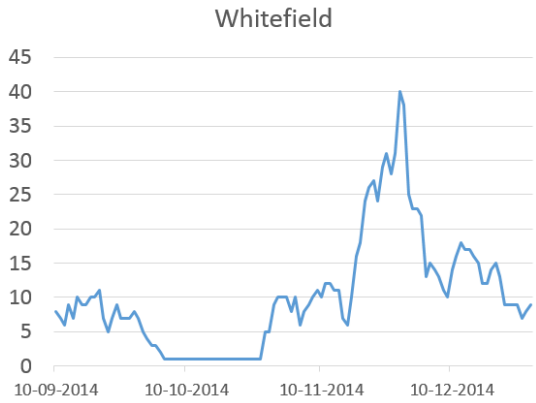

Figure 4: Spike Detected in location 'Whitefield' – Event detected 'Whitefield Rising' around '27-11-2014'.

For the seed event identification behind rise in discussions, we have analysed another location. Figure 5 displays the number of posts for the area 'ITPL, Bangalore' during same period as above. It can be seen that there is spike during the same duration and this location lies nearby to previous location. Proximity analysis of these locations helped us conclude that 'whitefield rising' event was seed for this problem.

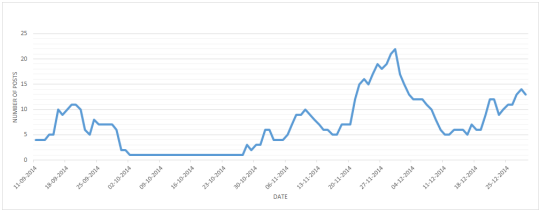

Figure 5: Spike Detected in location 'ITPL' – Impact of Event detected 'Whitefield Rising' around '27-11-2014'.

Table 1 describes the set of problems which occur frequently together. The prediction module, predicts the most severe problem with ≈ 76% accuracy. A division between actual and correct prediction for different problems for Bangalore city has been shown in Figure 6.

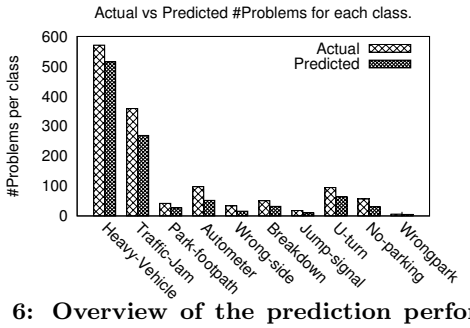

Figure 6: Overview of the prediction performance on given problem classes.

Year	Pattern
2014	{traffic_jam, heavy_vehicles} {Vehicle_breakdown , heavy_vehicles}
2013	{traffic_jam, no_parking} {traffic_jam, heavy_vehicles} {heavy_vehicles,Vehicle_breakdown }
2012	{traffic_jam, heavy_vehicles}
2011	{traffic_jam, heavy_vehicles} {No_parking,footpath_parking } {No_parking,auto_parking}

Table 1: Frequently co-occurring problems for last 4 years.

5. CONCLUSION

In this work we have presented an end to end system which captures social media data along with some external sources such as weather and news data, prepossessing to extract useful information, gain insights by analysing and make prediction using this data. We monitor trends in posts at a location to detect an event and return sentiment of people with respect to it. We use frequent pattern mining to identify commonly co-occurring problems. The system predicts severe problems for the next day using the insights from preceding days' data. Thus, this end to end system is a very essential tool for the traffic personnel to get useful insights of problems faced at different locations of the city and channel their actions accordingly. Further, data from different sources is clubbed together to come up with a better prediction. We plan to extend the problem prediction for fine grained prediction and incorporate more features like timeslot information etc. to make predictions for a time in a day as opposed to entire day.

6. REFERENCES

[1] E. J. Horvitz, J. Apacible, R. Sarin, and L. Liao. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. *arXiv preprint arXiv:1207.1352*, 2012.

[2] M. Ni, Q. He, and J. Gao. Using social media to predict traffic flow under special event conditions. In *TRB*, 2014.

[3] Wikipedia. Apriory algorithm. http://en.wikipedia.org/wiki/Apriori_algorithm.

[4] S. Yu and S. Kak. A survey of prediction using social media. *arXiv preprint arXiv:1203.1647*, 2012.