A City Traffic Dashboard using Social Network Data

Conference Paper · March 2015

DOI: 10.1145/2778865.2778873

CITATIONS
12

READS
507

4 authors, including:

Bidyut Patra
National Institute of Technology Rourkela
31 PUBLICATIONS **693** CITATIONS

SEE PROFILE

Abhishek Agarwal
Vermont State University
163 PUBLICATIONS **944** CITATIONS

SEE PROFILE

A City Traffic Dashboard based on Social Network Data

Apurva Pathak*
Microsoft R & D Pvt. Ltd.,
Hyderabad, India
appatha@microsoft.com

Abhishek Agarwal
Microsoft R & D Pvt. Ltd.,
Hyderabad, India
abagarwa@microsoft.com

Arnab Chakraborty
Microsoft R & D Pvt. Ltd.,
Hyderabad, India
arnab.chakraborty@microsoft.com

Bidyut Kr. Patra
NIT Rourkela, Rourkela, India.
patrabk@nitrkl.ac.in

ABSTRACT

With the growing urbanization and globalization, long commute and traffic problems have become the everyday nightmare of an Indian metro city dweller. The non-existence of a singular dashboard, which can provide holistic view of the city traffic, has aggravated this problem manifold for the traffic authorities and its citizens. The current state of the traffic monitoring, analyzing and comparison is sadly limited to tweets and posts of traffic citizens and do not provide enough actionable information. We propose to design a dashboard for obtaining real-time view of the traffic data scattered across various user status updates, tweets and comments on social networks. Proposed dashboard can provide a straight actionable information to the users and traffic authorities for handling traffic issues in efficient manner. The state-of-the-art machine learning algorithms are exploited for analyzing traffic data obtained from various sources.

Categories and Subject Descriptors

H.4 [**Information Search and Retrieval**]: Miscellaneous; D.2.8 [**Machine Learning**]: Metrics

General Terms

Theory

Keywords

Dashboard for traffic issues, Random Forest, Vowpal Wabbit Learning System

1. INTRODUCTION

Rapid urbanization is a good indicator of economic progress of a country. However, it has led to many socio-economic problems in India. Traffic scenario in Indian cities is one of the major problems and it becomes severe for densely populated cities like Bangalore, Chennai, Delhi and Kolkata. Government agencies have taken various measures for improving the traffic situation in these cities by improving public transport system, road infrastructures, etc. Recently, city traffic authorities have started feeding real-time traffic movements in text, image and video form into the online social networks such as Facebook and Twitter for its commuters. However, it does not provide actionable information for them. The need of this hour is a singular view, which provides a complete traffic state, which can be used by the traffic authorities and as well as the commuters for better management of traffic scenarios in those cities.

In this article, we propose to design a dashboard for obtaining real-time view of the traffic data scattered across various user status updates, tweets, comments on social networks, government domains, etc. We analyze this traffic information from different perspectives by exploiting state-of-the-art machine learning algorithms. We describe the proposal in detail in the next section.

2. ARCHITECTURE

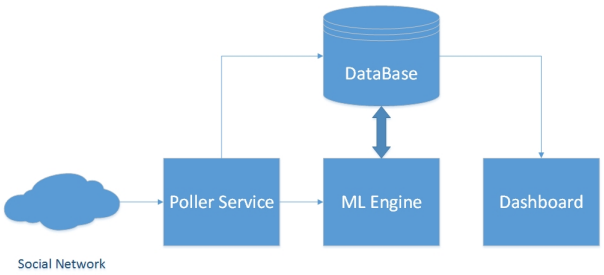

Figure 1: System Architecture Diagram

The architecture of the system is shown in Figure 1. A brief description of each of the module is given below.

- *Poller Service*: Polls the data source like Facebook and Twitter to get the latest feeds and stores it in the Database.
- *ML Engine*: It is invoked by the Poller Service after it puts the latest feeds in the database. ML Engine then applies text analysis techniques described in

*Corresponding Author

CODS 2015 2nd IKDD Conference on Data Sciences, Bangalore, India

4 and stores the processed information back into the database.

- *Dashboard*: A web interface to show the processed information stored in the database in an understandable way.

3. DATA SOURCE

The system uses data from various sources for analyzing traffic scenarios of Kolkata, Delhi, Chennai and Bangalore. Data sources used in this proposal are mentioned in Table 1.

Table 1: Data Source

Facebook	Bangalore Traffic Police
	Chennai Traffic Police
	Delhi Traffic Police
	Kolkata Traffic Police
Twitter	Traffline Bangalore
	Traffline Chennai
	Traffline Delhi
	Traffline Kolkata
Government	https://data.gov.in/

4. DATA ANALYSIS TECHNIQUES

We have used different data analysis techniques for Facebook and Twitter as we are extracting different information from them. The following section describes these techniques in detail.

4.1 Facebook

The data from the Facebook pages are mined and classified into one of these classes: Road Congestion, Road closure, Transportation refusal, Authority issues, Traffic rules violation, Other issues, Appreciate traffic police and Noise. The steps involved are as follows.

1. Create a training set having sufficient data for each class.
2. Preprocess posts: stemming, remove punctuation and stop words.
3. Obtain features from the posts using Vowpal Wabbit [2].
4. Classify posts using Random Forest [1].

As our system is real time the accuracy for this approach could not be calculated precisely. However, an accuracy of 93.25% has been obtained on 500 FB posts of Kolkata traffic Police with the training set size of 150.

4.2 Twitter

The tweets from Traffline are mined to extract the traffic (slow, moderate and fast) in various locations within a city. This information is then used to analyse the roads and localities with the heaviest traffic during a week. The steps involved are as follows:

1. Create a training set having sufficient data for each class (Relevant data, noise).
2. Preprocess posts: stemming, remove punctuation and stop words.
3. Obtain features from the posts using Vowpal Wabbit [2].
4. Classify posts using SVM and remove noise.
5. After removing the noisy posts we classify these posts into one of these classes: Slow, Moderate and Fast Traffic, using the method described in 4.1. We then apply Named Entity Recognition [4] to obtain the *Location* from the posts classified as Slow Traffic.

Just like facebook, classification accuracy of 98.46% was obtained on 500 tweets of Traffline Bangalore with training set size of 150.

4.3 Data.Gov.in

Unlike Facebook and Twitter, data from this source is formatted and hence, we directly consume data from these source without processing. The data includes the accident statistics (Total accidents, Fatal accidents, Killed and Injured) in Bangalore, Chennai, Delhi and Kolkata for the year 2008 to 2013. These data are shown and compared across various cities using line graphs.

5. IMPLEMENTATION DETAILS

The data from various social networks is polled periodically by a web service hosted in Microsoft Azure Web Role. Azure Machine Learning Studio [3] is then used for text analysis and processing 4. SQL Azure is used as database in our system. Finally the front-end web application is developed using ASP .NET and is hosted as Azure Websites.

6. DEMONSTRATION

A complete video demonstration of the dashboard can be found at http://youtu.be/3DqG_eYTl1o (1080p recommended).

7. CONCLUSION

We exploited state-of-the-art machine learning algorithms for analysing data obtained from different sources to provide a dashboard about the state of traffic in a metro city in India. We also provided a comparative scenarios among the cities. This work can be extended by incorporating weather information with the data obtained from difference sources and analysing them. Future traffic congestion can be handled using prediction techniques.

8. REFERENCES

[1] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] J. Langford, L. Li, and A. Strehl. Vowpal wabbit online learning project, 2007.

[3] Microsoft azure machine learning studio. `https://studio.azureml.net/`.

[4] E. F. Tjong Kim Sang and F. De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics, 2003.