Proposal for Summer Undergraduate Research Award (SURA)

Event Map Generator

Chinmay Rai
Computer Science & Engineering
(2016CS50615)
CGPA: 8.73
Contact : 826064810
cs5160615@iitd.ac.in

Samarth Aggarwal
Computer Science & Engineering
(2016CS10395)
CGPA: 9.57
Contact : 9810201131
cs1160395@iitd.ac.in

Prof.Mausam
Facilitator
Department of CSE
mausam@cse.iitd.ernet.in

Prof. S. Arun Kumar
Head of Department
Department of CSE
sak@cse.iitd.ernet.in

1 Introduction

An 'Event Map' is an association of the events happening in the world to their location of occurrence. For example, if we mark the locations of road accidents across India, what we will get is a special event map known as 'Crash Map'. If we mark the locations of all the crimes on the map, then we get another event map which we call as the 'Crime Map'.

Event maps have a wide variety of applications. Such visual descriptions of data prove to be very useful in drawing conclusions that are not apparent from textual data. The crash map of an area can help us identify the locations that are highly prone to accidents. With the knowledge of the location of the accidents, we will be better equipped to deal with them. For example, if the accident prone area happens to be a residential area, we may add more zebra crossing and subways to the area. On the other hand, if the area turns out to be one with blind turns, it is more effective to convert the roads to one-way traffic.

The crime map can help us identify the areas that are more prone to crimes. We then may deploy more cops and enhance patrolling in the area to keep the crimes under check. The event map of disease spreading can give us insights of the health care scenario in different regions. It will allow us to track the spread of different diseases and identifying the area of origin of a disease will also indicate the possible causes of a given disease. Pollution turns out to be major health issue these days. However, we still lack concrete facts regarding the contribution of different sources of pollution. An event map of respiratory abnormalities and air quality index of different locations can answer the above question. Also, if the area affected majorly turn out to be industrial area, we can conclude that industries are the major cause. If it turns out to be roads and highways, we can conclude the traffic to be the principal cause of pollution.

So location intelligence empowers us to analyze and take targeted measures to improve the situation in that particular area. The question that arises is that why are the available maps and data not sufficient to serve the cause. This is because most of the data that we have is in the form of text which is difficult to analyze without visual depictions. Moreover, the available maps provide data for city/state wide areas which fails at getting targeted information for the fore-mentioned purposes. Also, the organized data is somewhat outdated and loses relevance with its age. The recent data is rather unorganized and hence cannot be operated upon directly.

The essence of our project is to capture this need to collect and organize real time feed of data and come up with event maps for various utilities.

1.1 Related Work [1]

It has been established in various researches that significant reductions of accident impact can be achieved through effective detection methods and corresponding response strategies. As an essential component of traffic incident management, accurate and fast detection of traffic accidents are critical to modern transportation management. Countries like USA have have developed a detector networks on their automobile network which can monitor accurate location and time of accident, but a lot of maintenance and reliability issues associated with such a large scale network have forced the development of new techniques to monitor the accidents. In the Indian Scenario, apart from traffic accident, the daily traffic may also be obstructed due to reasons like political Possession, road construction etc. Electronic media can prove to be a useful source of information which can process to draw appropriate inferences Information from social media sources like Twitter is both noisy and unstructured. An effective text mining method is necessary to extract the useful accident-related information from tweets.

The Washington DC police Department has come up with a crime map [4] to track crime activities. This is the kind of map that we are aiming to generate in the Indian Scenario.

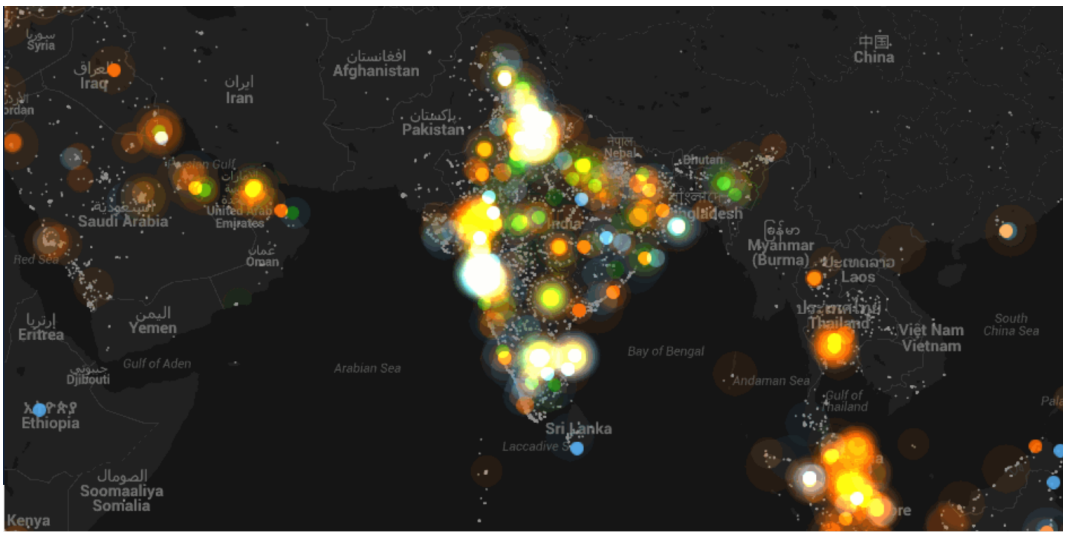

Figure 1: Event Map of India capturing support to different political parties (source - CartoDB)

2 Objective

Our aim is develop a tool that utilizes the live sources of data such as newspapers and twitter feeds [3] to generate event maps. We may as well use news archives of the recent past and other such less recent sources of data to increase the volume of data available. Next, we will apply Natural Language Processing tools to infer the data collected and organize it appropriately. After that we will apply machine learning tools to classify the data according to the nature of event map desired. This will depend on what kind of events do we want to get an event map for. Finally we will depict the results visually using visualization tools such as CartoDB.

3 Approach To Research And Development

The Event mapping task can be sequenced into the following set of steps.

1. Crawling Web Data Using Open source Web Crawlers.
2. Using Open Information Extraction Tools to determine the relation between Various clauses of a sentence.
3. We draw location,typical and characteristic inferences by forming clusters of data using a learning algorithm.
4. Validation of classification made our system by comparing it with a manually curated data set for which we will use crowd sourcing platform Amazon Mechanical Turk.
5. Visualization of the inferences drawn using location intelligence tool open Source CARTODB.

3.1 Web Crawling And Pre-processing

A web crawler (also known as a web spider or web robot) is a program or automated script which browses the World Wide Web in a methodical, automated manner. We Wish to use an open Source Web Crawler to extract textual information from electronic media sources. This Information is then processed into a number of standalone sentences which can be fed to the open IE system.

3.2 Open Information Extraction

The Open Information Extraction (Open IE) [2] system from the University of Washington (UW) and Indian Institute of Technology,Delhi (IIT Delhi) runs over sentences and creates extractions that represent relations in text.Open information extraction (open IE) refers to the extraction of

relation tuples, typically binary relations, from plain text. The central difference is that the schema for these relations does not need to be specified in advance; typically the relation name is just the text linking two arguments. For example, Barack Obama was born in Hawaii would create a triple (Barack Obama; was born in; Hawaii), corresponding to the open domain relation was-born-in(Barack-Obama, Hawaii).

3.3 Learning From Data

We classify each event based on location,type,characteristics and nature of its relation with other events in near time or distance.

3.4 Validation of Inferences Drawn

We will need human validation of classification and filtering done by learning algorithm in order to determine the accuracy of the predictions made To start, we two can be the two humans annotating the data we collect. But for the evaluation to be statistically valid, we might need a large enough text corpus to be manually annotated. Amazon Mechanical Turk (MTurk) is a crowd-sourcing Internet marketplace enabling individuals and businesses (known as Requester) to coordinate the use of human intelligence to perform tasks that computers are currently unable to do. We will use AMT to get annotated data to compare with the Data processed by our system.

3.5 Visualization

We use location Intelligence tool CARTODB which is an open source tool that allows for the storage and visualization of geo-spatial data on the web.

4 Tools And Resources

The Tools and resources which will be used during the course of project.

1. CartoDB.
2. Amazon Mechanical Turk(AMT)
3. Open IE 5.0
4. We will require access to use the super-computing facility (HPC) at IITD for processing large amount of data, which will be available from crawling various electronic media sources.
5. The Archives of major Indian Newspaper for the post 2000 era are available open source.(Hindu(http://www.thehindu.com/archive/); Times of India(https://timesofindia.indiatimes.com/archive.cms); Telegraph(https://www.telegraphindia.com//archives/)

5 Research Questions

We will try to address the following questions as a part of our project.

1. Is social media an appropriate source of data that gives adequate representation of different sections of the society in India?

2. Are the generated event maps biased/more comprehensive for metro cities or are they equally rich in data for the entire country? This will also cast light on whether this method of event map generation is reasonably useful for countries like India or will it be successful only in the American setting.

6 Budget And Duration

6.1 Budget

Amazon Mechanical Turk is a paid service, which will be used by us to annotated a corpus of data in order to validate our results. The exact amount of expenses will be clear only after we get quotation from AMT. However we expect the amount to be close to 25K.

6.2 Duration

The proposed duration work is two months, that is we expect to complete it by mid-july.

References

[1] *A Deep Learning Approach for Detecting Traffic Accidents from Social Media Data*, 2018 - by Zhenhua Zhang, Qing He, Jing Gao, Ming Ni.

[2] *https://github.com/dair-iitd/OpenIE-standalone* - Open Information Extraction (Open IE) system from the University of Washington (UW) and Indian Institute of Technology,Delhi (IIT Delhi)

[3] *https://homes.cs.washington.edu/ mausam/papers/kdd12.pdf* - Open Domain Event Extraction from Twitter

[4] *http://crimedc.com/* - Crime Map by DC police union