CLARIFYING SENSOR ANOMALIES USING SOCIAL NETWORK FEEDS

BY

PRASANNA GIRIDHAR

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Adviser:

Professor Tarek Abdelzaher

Abstract

The explosive growth in social networks that publish real-time content begs the question of whether their feeds can complement traditional sensors to achieve augmented sensing capabilities. One such capability is to *explain* anomalous sensor readings. Towards that end, in this work, we build an automated anomaly clarification service, called ClariSense. It explains sensor anomalies using social network feeds. Explanation goes beyond detection. When a sensor network detects anomalous conditions, our system automatically suggests hypotheses that explain the likely causes of the anomaly to a human by identifying unusual social network feeds that seem to be correlated with the sensor anomaly in time and in space. To evaluate this service, we use real-time data feeds from the California traffic system that shares vehicle count and traffic speed on major California highways at 5 minute intervals. When anomalies are detected, our system automatically diagnoses their root cause by correlating the anomaly with feeds on Twitter. The identified cause is then compared to official traffic and incident reports, showing a great correspondence with ground truth.

To My Parents and My Friends.

Acknowledgment

I would like to express my deep gratitude to Professor Tarek Abdelzaher, my research supervisor, for his patient guidance, enthusiastic encouragement and useful critiques of this research work. My grateful thanks are also extended to Md Tanvir Amin, Ph.D. student in Department of Computer Science, UIUC for his constant support in building our system, Lance Kaplan and Jemin George, US Army Research Laboratory, and Raghu Ganti, IBM Research, USA, for their help in offering me their valuable suggestions throughout in improvising the performance of our system. Finally, I wish to thank my parents for their support and encouragement throughout my study.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The proliferation of sensors utilized in human spaces, such as smart power meters, pollution meters, GPS devices, and vehicular traffic flow sensors, suggests that phenomena measured by such sensors will often be observed and reported socially as well. This is especially true of phenomena that deviate from the norm, hence attracting human attention. For example, the use of a mall's parking lot by freight trucks that increase local pollution, the closure of a freeway due to fire from a car crash, or the change in building occupancy patterns due to shutdown of a local employer are events that leave a signature on both local sensors and social media. This leads to the idea of developing a service that explains anomalies seen by sensors using data feeds from social networks (e.g., Twitter).

The general working principle of such a service is conceptually simple. Given a sensor network, such as the network of traffic flow meters on city highways, and given a social network, such as Twitter, the service detects (i) anomalies in sensor reports and (ii) anomalies in rates of different keywords on the social network. The two sets of anomalies are then matched up. A client of the service can hence see a map of anomaly locations (from the sensor network) and their explanations (from the social network) clarifying the corresponding likely root causes. This research work describes the design and implementation of such a service, called *ClariSense*. To our knowledge, this is the first attempt at jointly exploiting physical sensors and social feeds to both detect and explain anomalous events.

We evaluate our service by considering real-time data feeds from traffic flow sensors in three major cities in California; Los Angeles, San Francisco, and San Diego. Data was collected for 14 days from 4558 sensors at a period of one sample (per sensor) every 5 minutes. The data covered several anomalous events. Simultaneously, Twitter feeds were collected and ranked based on inclusion of "anomalous" keywords; specifically, keywords that occured with disproportionately different frequencies before and after the sensor anomaly. Anomalous sensor readings were then matched to anomalous tweets, offering clarifications of possible causes of the anomaly. A comparison of these explanations to manually collected ground truth suggests that ClariSense is very good at explaining truly unusual events. The more unusual the event, the better the chances that our service automatically explains it.

This work is broadly related to sensor network literature on self-diagnostics and health monitoring. Several services were developed in recent years that attempt to monitor the health of a sensor network and identify, in an automated fashion, root causes of anomalous behavior. While past work focused on detecting failures in the sensor network itself, the current research work searches for external physical events that can possibly explain the anomalous readings. It is therefore complementary in nature to past self-diagnostic efforts.

1.1 A Feasibility Argument

On June 9th, an anomalous 10-mile traffic jam was detected on a major Southern California freeway. To explain it, we contrasted two Twitter feeds; Namely, (i) Twitter feeds with keywords "California", "Traffic", and "Jam" shortly after the anomaly and (ii) Twitter feeds with the same keywords from one week away. The comparison returned the words "Impeach" and "Obama" as discriminative (in that they occur very frequently in the former feed while being absent from the latter).[1] Picking the most common tweets containing these keywords, we obtained the following actual tweets:

@OccupyOakland: "Impeach Obama" Gathering In California Causes 10 Mile Traffic Jam

@justpipertoo: Impeach Obama Rally Causes 10 Mile Traffic Jam on a Southern California overpass

Indeed, it turned out that the traffic jam was caused by a rally organized by the Tea Party. The above example clearly demonstrates the utility of exploiting correlated anomaly detection in social and sensor networks for explaining the causes of anomalies seen by sensors. While the above was done manually, the incident suggests that a general sensor anomaly *explanation* service can be developed that aims to automatically clarify root causes of sensor anomalies. The architecture, implementation, and evaluation of such a service are the topic of this research work.

In the rest of this research work, we use the California highway system as a running example. California logs traffic speed measurements on major routes and freeways every 5 minutes, together with related police reports, lane closures and other significant traffic events. A public database containing this information is available at: http://pems.dot.ca.gov. Simultaneously, we zoom-in on the most anomalous tweets about the freeway system, while removing isolated tweets and noise. By combining anomaly detection on Twitter and anomaly detection in the underlying traffic speed sensors, an anomaly explanation capability is attained, as described in the following.

[1] This example is presented purely for its illustrative quality of the capability described in this work, and with no intent for political innuendo whatsoever.

Chapter 2

System Design

The general architecture of our service is shown in Fig 2.1. As the figure suggests, several challenges need to be resolved for the service to work properly. We first describe a generic framework giving an outline of each of these challenges and then we present a detailed study focussing on a real world scenario, i.e., road traffic network.

We initialize the system with social feeds and data from physical sensors. These sensors are as diverse as microphone recording sound levels around a city, video cameras measuring optical flow, pollution sensors, etc. No matter the sensor, we would like to be able to determine if sensor reading is the result of a physical manifestation that can be discovered through passive crowd sourcing by scanning social media. The main contribution of this work is the design and implementation of a system which automates this scanning process to find related social feeds in time and space. However discovering anomalies in physical sensors remains out of the scope of this work and we use a black box to overcome this challenge. There have been several research works [17] carried out in the past for anomaly detection in sensors. The sensor anomaly detection in general on target count data is a two step process involving a learning stage and a likelihood based outlier detection stage. For the next step, our algorithm first determines the most discriminative feeds from the given set of social feeds. The main challenge in determining such feeds is to first identify the set of valuable keywords that uniquely characterize them and also occur at a higher proportion after the physical sensor event in comparsion to an earlier time. Considering the fact that texts used in social network feeds are mostly unstructured, taking advantage of a novel yet simple approach to overcome this challenge is the focus of this work. Looking simply at the occurance rate or percentage of increase in keywords before and after the events would more often produce noisy results and instead we use the concept of entropy to identify the amount of randomness introduced by the keywords to determine how much information can be gained from them after the event. Additionally the entropy value associated with the keywords allows us to rank the feeds ranging from most to least discriminative for the provided data set. The next task after the identification of discriminative feeds is to find the matches for physical events from the ranked list. This in itself can be an interesting challenge as the physical sensors only provide data and timestamp associated with the event. Searching for feeds based only on time would be of little use as this would again lead to noisy results even though discriminative but completely irrelevant to the sensor event. To overcome this challenge we use the spatial

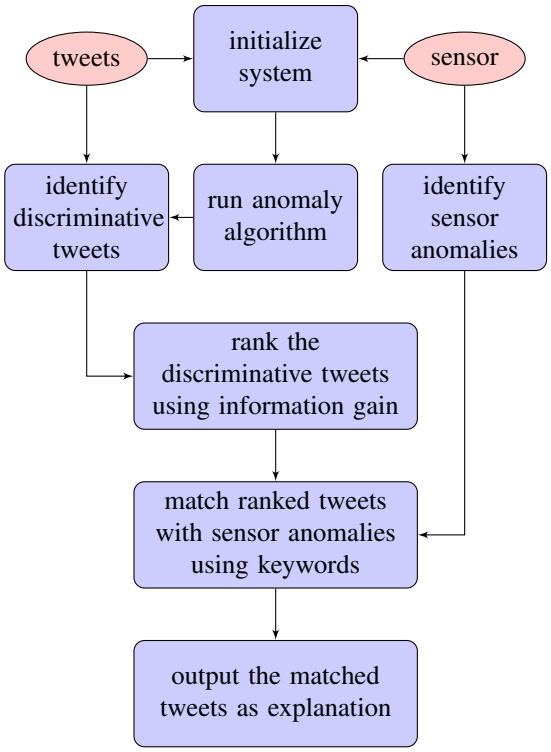


Figure 2.1: ClariSense Architectire

data associated with these sensors which are more likely to be used as keywords in the social feeds. Once we have identified the spatial data, the last step in our system is to find matches for sensor events providing a likely root cause. The mapping of sensor event to social feeds is then displayed to the user for verification against a ground truth.

In the following subsections we provide a detailed overview of the described challenges using Twitter as the social network and road traffic sensing devices measuring the speed, flow, occupancy, delay, etc. as the physical sensors.

2.1 Identifying Sensor Anomalies

The first challenge was to determine what constitutes a sensor anomaly. In many cases, the definition of anomalies is context sensitive. For example, what constitutes anomalous traffic at midnight might be perfectly normal at rush hour. We developed a simple algorithm that compares current sensor readings to “normal” readings in the same context. At present, we interpret context as day of the week and time of day. For the purpose of conducting an experiment to measure the performance of our system, the anomaly corresponding to flow equalling to zero by sensor readings during peak hours of the day (6A.M. to 10P.M.) was considered. The reason behind considering this as an evaluation case as opposed to other cases indicating anomaly like sudden deacrease in speed levels, increase in average delay

time, etc. was the higher probability of people tweeting online during an event of complete blockage of all the lanes of a freeway. We have also ensured in the sensor anomaly detection algorithm to indicate only blockages with a duration of minimum 45 minutes or more so that only major events get picked up. The choice of duration for a blockage can be tuned manually by the user before the processing begins. Thus sensor anomaly detection algorithm detects anomalies for all the highways in a given city and reports the start time, end time, duration and the sensor ID. It was also observed by manual verification that several anomalous sensor readings were not mutually independent and co-occured having an overlap of start and end time. Further it was observed that all such sensors were located on the same freeway number and direction as well. Thus in order to remove the redundant observations we have performed clustering of all such sensors using the distance between their geographical locations. The sensors with anomalous readings that were less than 2 miles in distance were clustered to denote a single event. This threshold of 2 miles was empirically decided by looking at how far a single anomally typically extended (event signature length distribution) for 5 major mainline blockages observed on highways in Los Angeles and the average length distribution computed using these blockages was 1.8599. Table 2.1 gives a brief overview of the event signature length distribution.

Table 2.1: Event Signature Length Distribution

Freeway	Duration (mins)	Extension length (miles)
5S	65	5.998
101S	60	1.5732
101N	75	0.258
10E	65	0.38733
405S	25	1.0825
Average		**1.8599**

2.2 Finding Discriminative Social Network Feeds

A long-running Twitter query continuously collects information that relates to traffic in California. Once a sensor anomaly is observed, we contrast the window of tweets ranked high during the anomaly (let us called it the current window) to a window collected and ranked before the anomaly (let us called it the normal window) to identify what changed in Twitter feeds. To zoom in on events that might be responsible, several challenges must be solved as described below.

2.2.1 Identifying Events

In order to identify anomalous events described in social network feeds, one must first identify what constitutes a distinct event. A first instinct might be to do some semantic analysis or language inspection of tweets and cluster together

those tweets that are about the same event. The problem with that approach is that it requires detailed language models and hence is hard to train and apply in situations where abbreviations, hashtags, and other slang are commonly used.

Instead, we opted for the minimalist approach. We hypothesized that when individuals describe the same event, they use overlapping sets of keywords. The question is: how large is the overlap, and how often does the overlap alone uniquely characterize an event? Since our running example uses a traffic flow sensor network, to answer the above question we collected tweets from Los Angeles, San Francisco, and San Diego that contain the word "traffic". We then clustered tweets collected on any given day by single keywords, keyword pairs, and keyword triplets, then inspected the clusters to answer two questions; (i) how many distinct events are described in tweets that fall into the same cluster, and (ii) how many distinct clusters describe the same event? Ideally, both answers should be close to 1 for a one-to-one mapping of events to clusters. Table 2.2 shows our results.

Table 2.2: Event Signatures

Signature	Events per Signature	Signatures per Event
Single Keyword	3.621	1.1579
Keyword Pair	1.1416	1.2725
Keyword Triplet	1.0628	0.4393

From the table, it is seen that keyword pairs come closest to a one-to-one correspondence between independent events and tweet clusters. We therefore adopt this approach in the rest of this work, identifying events by their keyword pair. Note that, there may be many ways to select the keyword pairs. Below, we describe how these pairs are selected.

2.2.2 Ranking by Information Gain

Our goal is to identify references to anomalous events. For the purposes of this work, we define such references as those that occur disproportionately frequently at the *current time*, compared to their average frequency. The "current time" is defined as a sufficiently large window, called the *current window* during which most tweets of the same physical event will likely arrive. For vehicular traffic problems, such as accidents, unusual traffic jams, fires, and closures, we empirically computed this window by obtaining the distribution of retweet delays (difference in time stamp between each retweet and the original tweet). We plotted a cumulative distribution as shown in Fig 2.2 where the X axis represents the delay in hours and Y axis represents the percentage of tweets for a given delay.

According to the figure, we set the current window to 24 hours as around 70% of events are retweeted within this delay period. It remains to define what "disproportionately frequently" means, when referring to references of an event in the current window. A few alternative definitions are possible.

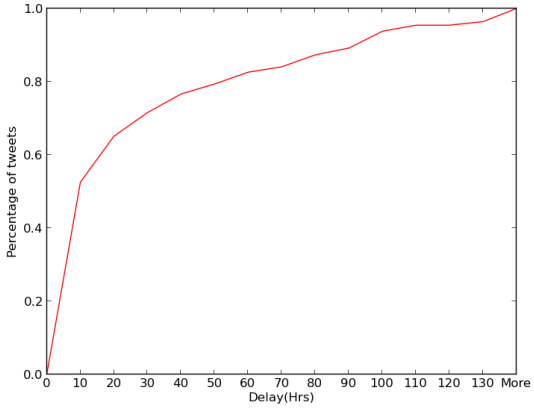


Figure 2.2: Distribution of retweet delays

The simplest way to identify disproportionately frequent event references in the current window is to simply compare the rate of these references (e.g., in references/day) within the current window to their average rate computed over a much larger past window. The problem with simply comparing the magnitudes of these rates is that it may favor high volume to disproprtionality. For example, a keyword pair such as (traffic, jam) that appears 50 times today, versus an average of 35 times a day, will have a larger rate difference compared to a pair such as (drunk, kills) that appears 12 times today compared to a past average of 0. Intuitively, the latter is a more anomalous event.

The above might suggest that we take the *ratio* of current to average rates, as opposed to the difference. This approach has its own problems. For example, a keyword pair that appears once today but never before will have a larger current-to-average rate ratio than a pair that appears 50 times today, but only once on average. Intuitively, the latter is more important. The former could be an entirely random occurrence (such as typo in one of today's tweets). An approach that combines both a large rate difference and a large rate ratio is therefore needed. One such approach, that has the advantage of having well-understood analytic semantics, is to use *information gain*. Let Y be a variable associated with the event type corresponding to normal ($y = 0$) or anomalous ($y = 1$) and X be a variable associated with absence ($x = 0$) or presence ($x = 1$) of a word-pair in the given window. The mathemtical underpinning of this approach are described below.

$$InformationGain = H(Y) - H(Y|X) \tag{2.1}$$

In equation 1 we find the entropy associated with variable Y(event type) and a conditional entropy of the variable

Y given another X(word-pair presence/absence). It is used to express how much extra information one still needs to supply on average to communicate Y given that X is already known. The mathematical formulae for computing these entropy values are defined using the following equations.

$$H(Y) = -\sum_{y \in Y} p(y) log_2 p(y) \tag{2.2a}$$

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x) \tag{2.2b}$$

$$= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) log_2 p(y|x) \tag{2.2c}$$

$$= -\sum_{x \in X} \sum_{y \in Y} p(x, y) log_2 p(y|x) \tag{2.2d}$$

In the above equations since the entropy of Y is constant for all words in the vocabulary, we need to evaluate only the conditional entropy of Y given X. Lower the value of this conditional entropy more discriminative the given word-pair is for that window and more will be its *Information Gain* value. The algorithm to find the discriminative Keyword Pairs using conditional entropy is described below.

Algorithm 1 Find discriminative events

Require: $n_X \geq 0, n_Y \geq 0$

1: $X \leftarrow tweets\_current\_window$
2: $Y \leftarrow tweets\_previous\_window$
3: $N_X, N_Y \leftarrow n_X, n_Y$
4: $V \leftarrow \{w_1, w_2, \ldots w_k\}\ \forall filter(w_i) = True$
5: **for** $pair_{ij} \in V$ **do**
6: $\quad count_{X_{ij}} = findCount(X, pair_{ij})$
7: $\quad count_{Y_{ij}} = findCount(Y, pair_{ij})$
8: $\quad Entropy_{pair_{ij}} = findEntropy(count_{X_{ij}}, count_{Y_{ij}}, N_X, N_Y)$
9: **end for**
10: $anomalousTweets \leftarrow tweets \in X\ \forall\ Entropy_{pair_{ij}} \leq 0.95$

Algorithm 1 analyzes tweets from current window and previous window to find the pair of words in the current window which give a conditional entropy below a *threshold*. Lines 1 and 2 represent the assignment of tweets from the two windows in variables X and Y. The length (number of tweets) of each window is assigned to variables N_X and N_Y in Line 3 and all unique "Noun" words occuring in each window such that they are not related to traffic in other types of network (mainly *internet*) are assigned to the variable V in Line 4. The filtering function *filter* performs part-of-speech tagging for all the unique words and returns false for internet related words or words that are not noun. The part-of- speech filtering was performed keeping in mind that events are described using nouns more often. From Line 5 to Line 8 the conditionaly entropy for each pair of words from the vocabulary is computed in an iterative way.

The $findCount$ function used in Lines 6 and 7 is used to find the count of a $pair_{ij}$ in a given tweet window set. Using the counts from each window the conditional entropy can be calculated using the function $findEntropy$ based on Equation 2d. Finally in Line 10 we return the set of tweets as anomalous which contain both the words in the discriminative pair along with the entropy value which can be used to rank the tweets. The threshold was set to 0.95 since it was covering 70% on an average of all the keyword pairs possible in the current window and it is very likely that a discriminative social feed if present will be selected. Also the range of the conditional entropy value for each pair is [0,1] and a value being more closer to 0 indicates that the pair is highly discriminative and gets ranked high. The threshold value can be manually tuned by user to limit the number of keyword pairs being considered and select only top k percent of tweets from the complete set of discriminative social feeds. We used a high value of threshold to analyze the complete set for our experiment.

2.2.3 Matching with Sensor Anomalies

In order to match the discriminative tweets with sensor anomalies we are using the geo-keywords associated with the landmarks around the physical location of sensors. With the help of google maps we are able to locate the sensors having anomalous behavior and also identify nearby buildings, streets, etc. that constitute our set of geo-keywords. These geo-keywords help us identify the potential explanations for sensor anomaly as they are more likely to occur in social feeds as compared to the sensor IDs or geographical coordinates associated with the sensors. We also take the advantage of using highway numbers and directions associated with these sensors as geo-keywords in order to increase the possibility of a match. These geo-keywords are then used together to form a query to find relevant tweets from the ranked list of anomalous tweets generated by Algorithm 1. The relevance at present is defined based on two factors: the first is match between geo-keywords of the query and the tweet and the second is the time stamp of the tweet. With the obvious notion that a tweet about an event will be reported only after its occurence we look for possibile explanations using the event report time and look not beyond a period of 24 hours from the start time. Thus based on these two factors we output the very first possible explanation along with a rank which is determined by the position of this explanation among all the discriminative tweets that can be identified using the geo-keywords. An approach described in [21] that addresses the problem of finding location information from unstructured text in social media can be further helpful for matching of discriminative tweets with the anomalous sensors.

Chapter 3

Implementation and Evaluation

3.1 Social Data Collection

We have performed experiments on three different cities from California, namely, Los Angeles San Francisco and San Diego for a period of two weeks starting from August 19, 2013 to September 1, 2013. The tweets were collected by a data crawling service using “traffic” as a keyword starting from the central location of a city and extending 30 miles in radius. The distribution of number of tweets collected on each day is presented in Figure 3.1 where Y-axis represents a date and the X-axis represents the number of tweets collected. Each city has a bar graph representation using different colors as indicated in the legend. As it can be observed from the figure that number of tweets collected for Los Angeles was insufficient on August 29, 2013 that can be attributed to a data collection glitch, we have eliminated that particular date and did not consider the sensor anomalies for evaluation which occured on that date.

3.2 Physical Sensor Data Collection

The sensor data corresponding to the three cities were retrieved from Caltrans Performance Measurement System (PeMS) [2]. This system provides data corresponding to flow values, average speed, occupancy and average delay for all major freeways. The physical sensors on these freeways record the values at every 5 minutes interval. It was observed that a particular sensor with ID 717697 located on 405 south bound freeway in Los Angeles was giving

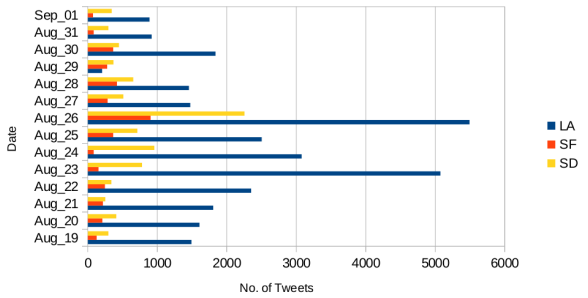


Figure 3.1: Distribution of tweets

zero flow values for a particular time duration for few consecutive days and neglected this particular sensor as it was a recurring event. A similar repeated occurrence of zero flow values was observed on highway 52 westbound for a sensor with ID 1114177 located in San Diego which was neglected as well. The sensor anomalies are not restriced to only those within Twitter query coverage of 30 miles but include all the anomalies for a given city.

3.3 Anomaly Detection and Evaluation

For each of the three cities we identified the sensor anomalies corresponding to flow equalling to zero on different types of freeways during peak traffic hours. The evaluation consists of two different measures. The first determines the number of explanations identified for sensor anomalies by Algorithm 1 and the second determines the average rank of these explanations. The first measure can be related to *Precision* which is defined as the fraction of explanations retrieved for all the blockages observed from the sensors. In order to evaluate the performance of our Information Gain algorithm we have used two other standard baseline methods to find discriminative tweets. The first baseline method picks out those tweets for which the pair of words have large difference between the current date and previous set of dates. The second baseline method picks out those tweets for which the pair of words have a significant increase in percentage between the current date and previous set of dates. A threshold of minimum difference of 3 for first baseline and a minimum percent increase of 50% for second baseline was set in accordance to the threshold set for Algorithm 1 for covering more than 70% of keyword pairs on an average and giving no advantage to our algorithm over the baseline methods. These two parameters can also be tuned by user to limit the number of tweets to top k per cent.

Table 3.1: Anomaly detection comparsion for different methods

City	Freeway Type	Total closures	IG	B1	B2
Los Angeles	Mainline	7	6	4	4
Los Angeles	Other	145	38	19	18
San Francisco	Mainline	40	21	7	8
San Diego	Mainline	16	8	2	1
San Diego	Other	14	2	1	1

Table 3.1 shows the comparsion between Information Gain algorithm with the two baseline methods for all the three cities. The anlysis has been presented seperately for *Mainline* since these are multilaned freeways with high volume of traffic at any given time as compared to other freeway types like off-ramps, on-ramps, freeway to freeway intersection that have less number of lanes (mostly limited to one) with a comparatively limited volume of traffic. A mainline blockage can be generally regarded as a major event occurence causing the traffic to a complete halt. It can be observed from the table that mainline blockages have been identified at an accuracy of 6 out of 7 using Information

Gain (IG) for Los Angeles and over 50% for San Francisco and San Diego. The low precision for these two cities compared to Los Angeles can be attributed to more number of tweets collected during collection for the latter city as seen in the Figure 3.1. The other freeway type blockages for Los Angeles has 111 off-ramp (FR) and on-ramp (OR) out of 145 observed blockages while San Diego has 10 on-ramp (OR) out of 14 observed blockages. The FR and OR blockages are mostly for single lanes with a really low volume of traffic. It can be expected that people tweet for such blockages very rarely as they do not occur due to major events which explains the low precision for these blockages in comparison to mainline blockages.

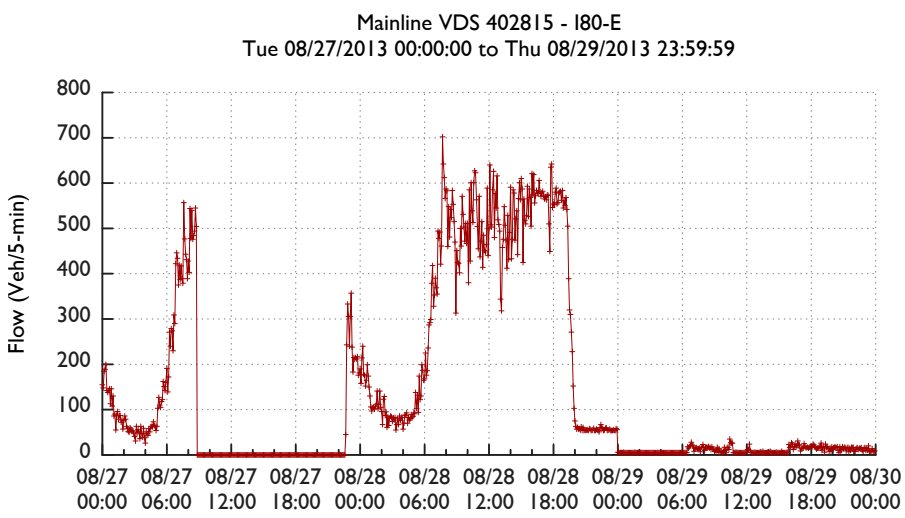
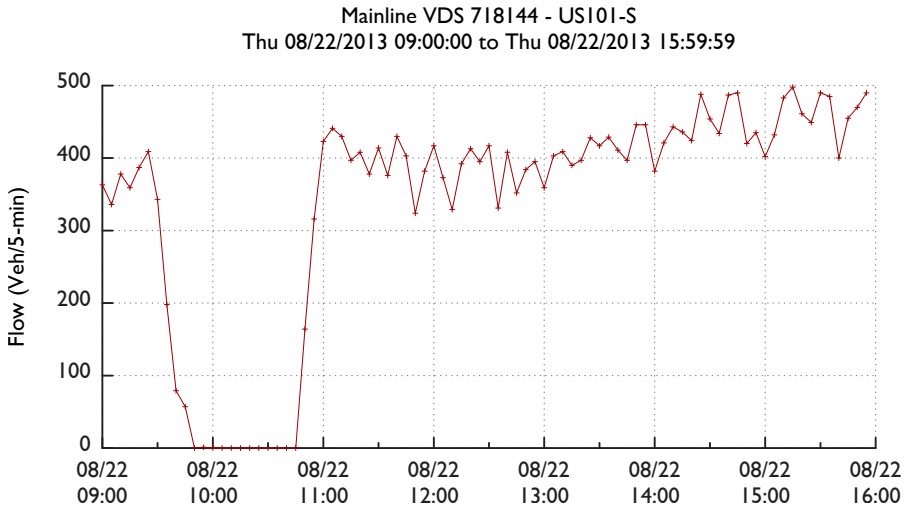
The second measure can be related to *Mean Rank* which is defined as the amount of lookup needed from top on average to get the explanations for an anomaly corresponding to a particular highway using geo-keywords associated with the sensor. To evaluate the performance of our Information Gain algorithm we have used a standard lookup method with no ranking scheme using geo-keywords related to highway to find discriminative tweets. Table 3.2 shows the average position of the tweets from the top using these two lookup methods. It was observed that on an average the relevant discriminative feed was found within top 2 positions.

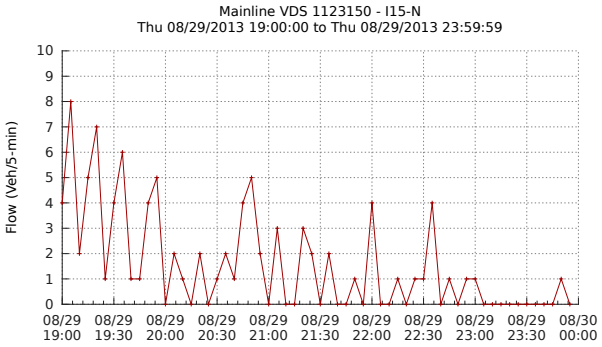
Table 3.2: Average position of tweets from top

City	Information Gain	Without Information Gain
Los Angeles	1.884615	22.76
San Francicso	1.30434	5.8695
San Diego	1	4.5

Some of the interesting mainline blockages captured by Algorithm 1 over the period of two weeks are indicated in Table 3.3. The table also includes a graph indicating the flow at every 5 minutes interval retrieved from PeMS [2] portal for one event from each city. The first of the interesting events captured is the Bay Bridge closure in San Francisco city which spanned over 5 days as indicated in the graph. Physical sensor anomalies were observed during this duration on few sensors located on I-80 highway and using their location coordinates we identified “Bay Bridge” as the geo-keyword associated with them. A lot of discriminative feeds were identified by our system relating to this event and two such have been indicated in the table. The next interesting event occurred in Los Angeles city on August 22, 2013 when all the lanes of US101 northbound and southbound were closed for a duration of around 2 hours. Our system returned a possible underlying cause for this blockage indicating the presence of bomb squad and police on the highway. On verifying from CHP incidents report it was concluded that this event indeed occurred due to the cause explained by our system. Another interesting event occurred in San Diego on August 29, 2013 on the I-15 northbound when several blockages were observed between 19:00 and 23:59 hrs. Our system returned an explanation relating to

Table 3.3: Interesting events captured by Algorithm 1 for zero flow blockages

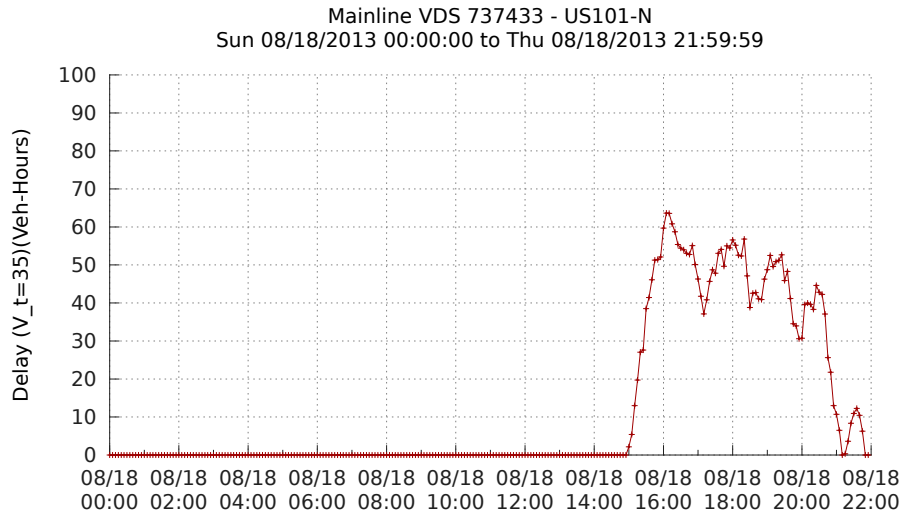
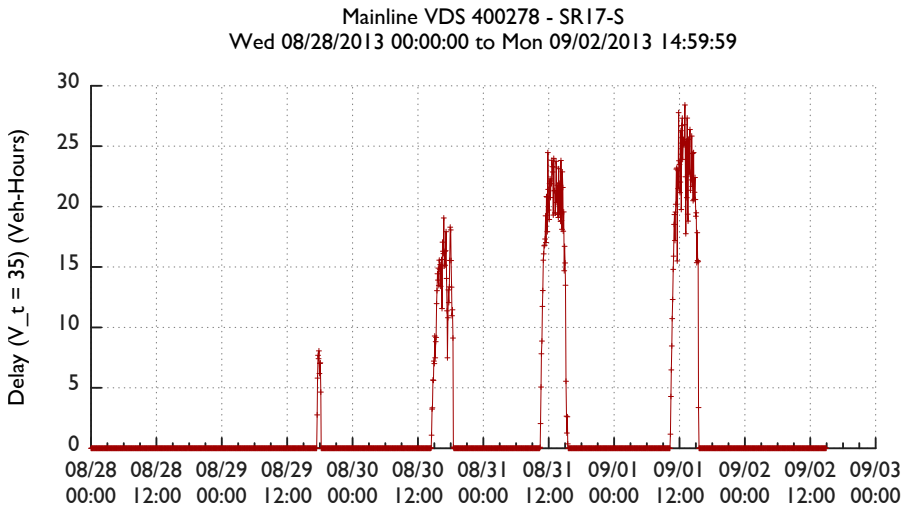
Date	City	Freeway	Keywords	Tweet Explanation
				
Aug 31	SF	80E	**Bay Bridge**	**Bay Bridge**: Day 3 construction to finish traffic switch-over proceeds without a hitch
Aug 29	SF	80E	**Bay Bridge**	Avoid the traffic caused by **BayBridge** closure & work-from-home! With SabaCloud you can
				
Aug 22	LA	101S	**Ventura Fwy**, Woodland Hills	Bomb squad & bunch of cops on the 101 **Ventura freeway** making more of a traffic. Good way to start the day *Rolls eyes*
Aug 22	LA	101N	**Ventura Fwy**, Woodland Hills	Update: All Lanes of the 101 Freeway Have been closed See Live Traffic Map Here - Police & Fire - Calabasas, CA Patch

Date	City	Freeway	Keywords	Tweet Explanation
Sep 1	LA	110N	LA Memorial Coliseum, **ML King Jr Blvd**	TRAFFICALERT: All lanes of NB 110 closed at **MLK** after pedestrian walking in lanes fatally struck by car; traffic diverted off at Vernon
				
Aug 29	SD	15N	Escondido, Avocado Hwvy	TRAFFIC ON **I-15** chargers: # SFvsSD Game Time Forecast: Kickoff (7p): 81 FHalf Time (8:30p): 73 FEnd of Game (10pm): 72 F with WNW 4 mph.

the chargers game played that night at the Qualcomm Stadium near Escondido Highway on I-15.

In addition to the detailed study over the two weeks period for our experiment and evaluations using the sensor anomaly corresponding to zero flow (complete blockage), we have also collected sensor anomalies relating to sudden spikes in delay or increase in occupancy for these three cities and present some interesting events in Table 3.4. However we do not present any evaluation measures like *precision* and *mean rank* for these anomalies with the aim to indicate the robustness of our system in explaining causes for other types of anomalies as well. The first of these is a spike observed in delay by a sensor located on US101 northbound in Los Angeles on August 18, 2013. The reason behind this anomaly was identified by our system as a small brushfire incident which occurred nearby that sensor. Another interesting event is a sudden increase in the delay observed by a sensor located on SR-17 southbound in San Francisco on September 1, 2013. There was a lot of traffic delay observed in this area due to the "Labor Day" weekend and our system returned a feed corresponding to this delay.

Table 3.4: Explanation for events related to other anomaly types

Date	City	Freeway	Keywords	Tweet Explanation
				
Aug 18	LA	101N	**Calabasas**	If you're wondering what's up with the 101 traffic & all those helicopters: there's a small brushfire in **Calabasas**.
				
Sep 1	SF	17S	**Campbell**	Bad beach traffic on 17 SB - starts around Hamilton in **Campbell** & cars creeping at under 10 mph. Steer clear!

Chapter 4

Conclusions

4.1 State of the Art

Although social sensing is an active area of research, the challenges of designing a robust and automated systems by exploiting the social network in the context of pervasive services and applications is still at an early stage. Modern cities are already provided of cameras and microphones for security purposes, and sensor networks for environmental monitoring. RFID-based tickets and badges can keep track of user movements and activities. Smart phones are embedded with gyroscope, compass, accelerometer, proximity sensors, and localization tools. With the help of a social sensing system we can enrich the scenario, by complementing the available information and thus enabling higher levels of context-awareness. For example, Rosi et al. have surveyed the possible approaches to such an integration, and discuss the open issues and challenges facing researchers [25]. The SensorFly project [23] develops a sensor cloud, which consists of many low cost and individually limited mobile sensing devices that only when functioning together can produce an intelligent cloud, in disaster situations such as an earthquake and fire. In [24] the authors have addressed the problem to detect a target event by devising a classifier of tweets based on features such as the keyword in a tweet, the number of words, and their context.

All past research on participatory sensing describes how to aggregate and clean-up collected data to present an event. For instance, the Nericell project [22] presents a system that performs rich sensing using smartphones that users carry with them in normal course, to monitor road and traffic conditions. The GreenGPS system [19] provides a service that computes fuel-efficient routes for vehicles between arbitrary end-points, by exploiting vehicular sensor measurements available through the OnBoard Diagnostic (OBD-II) interface of the car and GPS sensors on smart phones. This work complements that past work by looking into the underlying cause of the event. In many cases, the underlying cause is known via context, i.e., the volcano erupted. In the more general setting we need to find the cause via social sensing.

Especially, thanks to fast development of smartphones and social networks, participatory sensing is paid more attention in disaster response applications in recent years. People share their information about the disaster region to

social networks and special-purpose services, to help each other beat the disaster together. For instance, popular social networks such as Facebook [6] and Twitter [7], played an important role after natural disasters such as Japan Tsunami in 2011 [13] and USA Hurricane Sandy in 2012 [14]. Many service providers, some notable names including Waze [3] and GasBuddy [4], set up special-purposes services to allow people to participate and report the availability of gas stations after Sandy via web or smartphones.

A variety of applications can be created to collect real time information from large groups of individuals in order to harness the wisdom of crowds in a variety of decision processes. For example, the now phased out the Google Latitude [9] application collected mobile position data of uses, and used this in order to detect the proximity of users with their friends. This can lead to significant events of interest. For example, proximity alerts may be triggered when two linked users are within geographical proximity of one another. This may itself trigger changes in the user-behavior patterns, and therefore the corresponding sensor values. This is generally true of many applications, the data on one sensor can influence data in the other sensors. Numerous other GPS-enabled applications such as City sense [10], Macrosense [11], and Wikitude [12] serve as gps-based social aggregators for making a variety of personalized recommendations. The approach has even been used for real-time grocery bargain hunting with the LiveCompare system [18]. In contrast, our work takes advantage of the social networks, and has the ability to explore the sensing capability of the social networks and predict the underlying causes of events observed by the sensors in physical world.

4.2 Discussion and Future Work

We have shown thus far that building a service that can detect anomalies in the rates of different keywords on the social network and then map the anomalies in physical sensors with their explanations from the social network clarifying their likely root causes is indeed possible. In this Section, we will discuss the implications of our hypothesis and the experiments that we performed.

First and foremost, would a simple correlation between two time series, one from the Twitter feed and the other from physical sensor feed be enough to solve this problem? With the huge amount of data collected from Twitter which is constantly populated with real time updates, filtering out tweets based only on time and space would rather prove to be an expensive process for finding the correlation between the two series since there may be enough irrelevant tweets introducing noise. In particular, we are interested in finding tweets which generate valuable information to help identify the root causes of physical sensor anomalies. Hence, we need new algorithms to address this problem and once such is presented in our work for capturing the discriminative social feeds. Based on the results in this work,

we believe that finding such feeds is indeed possible. Our immediate future work is to determine the credibility of these discriminative feeds. Finding credibility in assured social sensing [20] has already been addressed in the past. At present, our system focuses mainly on finding discriminative feeds irrespective of the truth factor associated with them. Another direction is to take advantage of an earlier work [21] for better localization of tweets based upon the content in order to improve the precision. We plan to enhance the performance of our algorithm with the inclusion of truth factor and localization of tweets.

4.3 Conclusions

In this work, we have addressed the hypothesis of - is it possible to explain the likely causes of the anomaly to a human by identifying unusual social network feeds that seem to be correlated with the sensor anomaly in time and in space? We collected Twitter data over a period of two weeks for three different cities in California and simultaneously obtained log traffic records of the California Highway system from PeMS [2] public database. We first built a system to identify discriminative social feeds using the Information Gain algorithm and also identified anomalies in traffic log records using a sensor anomaly detection algorithm for traffic blockages observed during this period. Next we ranked these feeds based on entropy parameter and matched them with sensor anomalies. To evaluate the performance of our algorithm, we considered two baseline methods to find such discriminative tweets and observed that an accuracy of at least 50% in worst case was achieved by our algorithm as opposed to the baseline methods for the *Mainline* blockages. To a large extent, the accuracy depends on the amount of social feeds collected for a city on a particular day when anomalous incidents occur and also the volume of traffic flowing through a particular freeway. A high traffic volume indicates more people posting feeds on social network about an incident thereby increasing the discriminative factor of the feeds about an incident in comparison to other normal feeds. We also observed that the mean rank of lookup for a discriminative feed using our algorithm was around 2 on average in comparison to a standard lookup method. Thus we can expect the possible explanation for a sensor anomaly to be among the top 2 tweets in most of the cases. We concluded this work by discussing various directions that we are currently pursuing to enchance the performance of our system.

References

[1] The Google Galaxy Nexus phone. URL: http://www.google.com/nexus/#/tech-specs.

[2] Caltrans Performance Measurement System (PeMS) http://pems.dot.ca.gov/

[3] Waze: Free GPS Navigation with Turn by Turn. URL: www.waze.com/.

[4] GasBuddy: Find Low Gas Prices in the USA and Canada. URL: www.gasbuddy.com/.

[5] All Hazards Consortium. URL: www.ahcusa.org/.

[6] Facebook. URL: www.facebook.com/.

[7] Twitter. URL: www.twitter.com/.

[8] Google map. URL: maps.google.com/.

[9] Google Latitude. URL: latitude.google.com/.

[10] CitySense. URL: www.citysense.com.

[11] MacroSense. URL: www.sensenetworks.com/products/macrosense-technology-platform/.

[12] Wikitude. URL: www.wikitude.com/.

[13] Japan considers using social networks in disaster situations. URL: http://www.engadget.com/2012/08/30/japan-considers-using-social-networks-in-disaster-situations/.

[14] Twitter / Search - *#hurricanesandy*. URL: https://twitter.com/search?q=%23hurricanesandy.

[15] Gasoline Runs Short, Adding Woes to Storm Recovery, NY Times. URL: http://www.nytimes.com/2012/11/02/nyregion/gasoline-shortages-disrupting-recovery-from-hurricane.html?pagewanted=all&_r=0.

[16] GasBuddy Fuel Shortage Tracker. URL: http://www.gasbuddy.com/sandy/.

[17] Yuan Yao, Abhishek Sharma, Leana Golubchik, Ramesh Govindan Online anomaly detection for sensor systems: A simple and efficient approach Performance Evaluation, Volume 67, Issue 11, November 2010, Pages 1059-1075, ISSN 0166-5316

[18] L. Deng, L. P. Cox. Livecompare: grocery bargain hunting through participatory sensing. HotMobile, 2009

[19] R. Ganti, N. Pham, H. Ahmadi, S. Nangia, and T. Abdelzaher. GreenGPS: A Participatory Sensing Fuel-Efficient Maps Application. In Proceedings of ACM 8th Annual International Conference on Mobile Systems, Applications and Services (MobiSys), 2010.

[20] Dong Wang, Lance Kaplan, Tarek Abdelzaher and Charu C. Aggarwal On Credibility Estimation Tradeoffs in Assured Social Sensing IEEE Journal On Selected Areas in Communication (JSAC) Vol 31. No. 6 June, 2013

[21] Kisung Lee, Raghu Ganti, Mudhakar Srivatsa, Prasant Mohapatra Spatio-Temporal Provenance: Identifying Location Information from Unstructured Text In IQ2S workshop, co-held with PerCom 2013.

[22] P. Mohan, V. N. Padmanabhan, and R. Ramjee. Nericell: using mobile smartphones for rich monitoring of road and traffic conditions. In Proc. of ACM Conference on Embedded Networked Sensor Systems (SenSys), 2008.

[23] A. Purohit, Z. Sun, F. Mokaya, and P. Zhang. SensorFly: Controlled-mobile sensing platform for indoor emergency response applications. In Proccedings of 10th Information Processing in Sensor Networks (IPSN), 2011.

[24] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors In Proceedings of the 19th international conference on World wide web (WWW '10). ACM, New York, NY, USA

[25] Alberto Rosi, Simon Dobson, Marco Mamei, Graeme Stevenson, Juan Ye, Franco Zambonelli Social Sensors and Pervasive Services: Approaches and Perspectives Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on