Detection, Localization and Characterization of Transient, Urban Events using Multi-Modal Information

Kasthuri Jayarajah
Singapore Management University
kasthurij.2014@phdis.smu.edu.sg

Noel Athaide
Singapore Management University
noelathaide@smu.edu.sg

Vigneshwaran Subbaraju
Agency for Science Technology and Research (A*STAR)
vigneshwaran_subbaraju@sbic.a-star.edu.sg

Archan Misra
Singapore Management University
archanm@smu.edu.sg

1 PROBLEM STATEMENT

With increased attention on developing technologies for smarter cities, we see increased use of data from a disparate multi-modal sensors that are being deployed as part of the urban infrastructure. For eg. the Call Detail Records (CDR) from telcos enable continuous tracking of population mobility. Other examples include cameras mounted along roads (that help monitor traffic flow) and buses equipped with location sensors. Further still, users of social media platforms such as Twitter and LBSNs (e.g., Foursquare) serve as distributed social sensors, voluntarily sharing content related to events that occur in their localities.

In this work, we investigate a number of key questions: (1) first, can such multimodal sensors be used in detecting urban events of different categories and scale (e.g., a large musical concert vs. a small gathering), (2) second, are the disparate sources equally capable of localizing such events, both spatially and temporally, and (3) finally, could user-shared content such as text be used to semantically annotate such events.

We demonstrate the feasibility of detecting events through a set of sampled urban events and share early insights on the differences across the sources. Our preliminary findings show that CDR and bus availability data are able to detect at least 40% of the events within 1 km from the event venue, and that hashtags extracted from Twitter include keywords related to ongoing events. We also discover that both physical and social sensors show better detection during hours prior to the start of the event. We developed a web application (see Figure 1) that mines multimodal information, characterizes mobility, detects and displays potential events along with semantics assimilated from social media.

2 DATA SOURCES

In developing *EventXplore*, we leverage mobility information from the following data sources.

DataSpark *discretevisit* and *staypoint* API [1]: This API is used to extract the residency information of people at various locations based on their Call Detail Records (CDR). The discretevisit API provides the number of people entering a zone at hourly intervals throughout the day. In this study, we obtained data from 85 subzones of Singapore, which constitutes the greater CBD area where much of the events took place during the 2 month observation period (May and June 2017). Similarly, we use the staypoint API for ascertaining the number of people who stayed in a particular subzone for atleast 20 minutes.

[1]https://datasparkanalytics.com

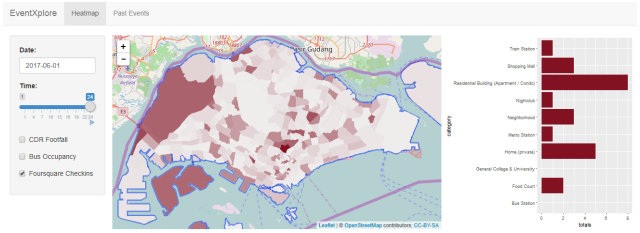

Figure 1: Landing Page View of the Web Application at: http://ares.smu.edu.sg/~kasthurij/videos/cikm-video.mp4

Transport data from LTA Data Mall [2]: We use the Bus Arrival API that provides the estimated time of arrival (ETA), and occupancy information for the 'Next bus' and 'Subsequent bus' for around 5000 bus stops island-wide. Similar to the case of CDR, we specifically sampled data from 162 bus stops serviced by 5 specific routes that passed through the greater CBD area. The geo-coordinates of the bus stops were also gathered using the API.

Public Tweets [3]: Tweets posted by users identified as being from Singapore and were visible publicly were crawled for the same period. Information such as the post ID, time of post, the tweet text, hashtags used, if any, and the geo-coordinates if the post was geo-enabled were extracted for each Tweet.

FourSquare Checkin data [4]: Based on publicly available Tweets, a subset of which were posts related to Foursquare, were also extracted along with the Foursquare venue, it's coordinates and the type of the venue (e.g., train station, shopping mall, etc).

3 ANALYSIS AND MODELING METHODOLOGY

In the design and implementation of our framework, we focus on the following key components illustrated in Figure 2.

Data Acquisition Layer: The API requests (e.g. DataSpark and DataMall) and crawler scripts (e.g., Tweets and Foursquare check-ins) assimilated the various attributes within stipulated query limits and temporal granularity. The raw data was primarily stored in ElasticSearch[3] for data sources with large volumes and frequent updates (e.g., every minute for bus data). **Data Processing Layer:** For each source, the aggregated occupancy level $c_{s,w,d}$ at location l, during window w, on day type d was summarized. For example, for CDR data, the set of all s were the 85 subzones whereas for bus

[2]https://www.mytransport.sg/content/mytransport/home/dataMall.html
[3]https://developer.twitter.com
[4]https://developer.foursquare.com

data, this was the collection of bus stops. The window w was hourly in the case of CDR, whereas we considered bins of 15 minute and half an hour lengths for the remaining sources. The day type was used to differentiate between weekdays and weekends. We assume that $c_{s,w,d}$ follows a normal distribution.

Event Analytics Layer: We considered a number of distance measures including Euclidean (bus data), z-score (checkins) and distance to the median (CDR), for declaring a sample $c_{s,w,d}$ to be anomalous. For combinations of sources, we fuse the outlier scores across the sources as the scaled, arithmetic mean. For time bins and locations that the system declares as outliers, we mine the types of Foursquare venues that received the most number of checkins, and hashtags from Twitter that had the highest $TF - IDF$ scores for annotating the anomalies.

4 ACCURACY VALIDATION AND METHODOLOGY

In this section, we describe our evaluation methodology and share our preliminary findings.

4.1 Ground-truth Events

To validate our approach, we looked up the web for events that happened in the Singapore, in the months of May and June 2017. Based on our search, we selected a few large scale, small scale and medium scale events tabulated in Table 1. The first 8 events in this table fall on a weekend, the next three fall on weekdays and the remaining are multi-day events that encompass both weekdays and weekends. During the period under consideration (May/June 2017), there were 42 weekdays and 19 weekends/public holidays. The public holidays fell on 1 May, 10 May and 25 Jun. In the evaluation in Section 4.3, we focus on a subset of "localized" events that were confined in locations and durations – for example, holidays were removed as they typically cause multiple, localized events throughout the day. For multi-day events, we focus on the detection of the start time on the first day of the event.

4.2 Event Detection with CDR

We first evaluated the ability of CDR data (DataSpark API) to localize events in space and time. The analysis was restricted to the 89 subzones we had earlier decided to focus on. We aggregated the discrete-visit and staypoint counts in each subzone and day into 5 time-bins (00:00-06:00AM, 07:00-10:00AM (AM-Peak), 11:00AM-05:00PM (Off-Peak), 06:00-08:00PM (PM-Peak), 09:00-11:00PM). These time-bins reflect the peak and off-peak activity of people in the city. The median, first quartile, third quartile, inter-quartile range and z-score values were obtained for the weekdays and weekends separately, for each subzone. If a particular value for a given time-bin, day and subzone was away from the first or third quartile by more than 1.5*times the IQR, it was declared an outlier. This way, a total of 864 outliers were identified across 42 weekdays, 5 time-bin and 85 subzones. Among these, it was also found that 747 were in the month of May, especially in the first two weeks of May. The outliers were also overwhelmingly on the lower side showing diminished flow of people across many of the zones. This may be attributed to the holiday period (1 May and 10 May being public holidays) when several of the residents could have taken a break

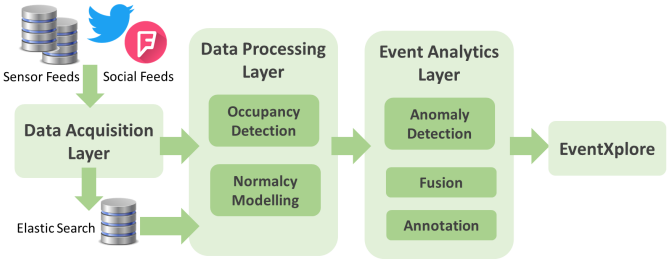

Figure 2: System Overview.

from their regular life. A similar effect could be found around the public holiday on June 25. Thus this simple outlier analysis can give an overview of the large-scale disruptions in the city. While the dates of the public holiday are known and people may expect to see an anomaly in the period, the extension of the anomalies over time and its characteristic distribution in space may not be obvious and it is brought out in a data driven manner by this simple approach. However, this approach may not be sufficient for smaller scale events. To explore such small scale disturbances, we next look at the *hourly* CDR data and more importantly, explore the use of multi-modal sensing (using bus arrival times and social media data) to achieve better recall of such fine-grained events.

4.3 Spatio-temporal variations due to sources

In this section, we seek answers for two key questions:

(1) *Spatial localization* - Does the event detection capability and the accuracy to which the event can be localized vary across the multiple sources?
(2) *Temporal localization* - Are the sources capable of detecting the start time of an event earlier than it occurs?

We measure the detection accuracy in terms of recall (i.e., proportion of events detected out of the known events in Table 1. We declare that an event is recalled by a source if the following criteria is met: (1) a location s is an outlier for the day type d and window w corresponding to the event date and time, and (2) s is within a radius R from the event venue. To understand the temporal bias, we vary window w as the same hour as the event start time, and an hour prior to that. We vary R between 0 and 4000 meters. In Figure 3, we plot the distance threshold R on $x-$axis and the recall on the $y-$axis for each source, for detection during the event start hour and the hour prior to the start, respectively. We make the following observations:

(1) In both cases, we observe that the physical sensors, i.e., CDR and bus, are better at detecting the events. Nearly 40% of the events were detected with a localization error of less than 1 km.
(2) All three sensors show predictive capability with the events being detected an hour earlier than the scheduled start time – however, we note that the recall is significantly better for the physical sensors.

4.4 Validation on Other Regions

To further validate the effectiveness of using mobility signals for detecting events, we accrued a separate dataset belonging to another region, from a different time period; we collected trip data

Table 1: Canonical set of events in Singapore in May/June '17

Date	Time	Name	Scale
10 May	All day	Vesak Day 3 step 1 bow procession	Large
3-4 Jun	PM peak	Dragon Boat Festival	Medium
10 Jun	All day	Ultra Singapore Electronic Music Festival	Medium
17-Jun	PM peak	Bark and Kisses: A dog cafe adventure	Small
17-18 Jun	PM peak	Urban Camping	Small
17 Jun	All day	Food Expo	Medium
24-Jun	All day	Hari Raya Market	Large
24-Jun	Off-peak	Dreamworks day	Small
14-Jun	All day	Natl. Inter-School Dragon Boat Championships	Medium
16-Jun	PM peak	ADAC 2017 music concert	Small
30-Jun	PM peak	Britney Spears music concert	Large
16-17Jun	PM peak	OMM:Hensel and Gretel	Medium
1-4 Jun	All day	Singapore Intl. Piano Festival	Small
9-11 Jun	All day	Health Fiesta	Small
9 Jun	PM Peak	A-MEI<Utopia 2.0 Carnival> World Tour	Medium
9-11 Jun	All day	Doctors without borders, Sg. Intl. Film Festival	Small

(pickup/dropoff location and pickup/dropoff time) of Yellow taxi-cab trips that started and terminated within Manhattan during the whole year of 2013[5] and venue check-ins from Foursquare for the same period. A total of 143 million taxi trips and 24 million check-ins across 24,990 venues were analyzed. The taxi pickup and dropoff locations and the Foursquare venues were aggregated spatially to the Census Tracts[6]. In addition, we manually labeled the location coordinates and start/end times of events during the period across Manhattan for a list of 160 events based on NYC Insider Guide[7]. We repeated the analysis of recall performance on 69 of those events as we discarded events whose location and time were unclear or those that spanned multiple routes/blocks such as in the case of parades (e.g., Macy's parade).

In Figure 4, we plot the recall performance for (1) the event start hour and (2) three hours after the event started (presumably closer to the end of the event), for the three signals: taxi drop-offs (blue solid line), taxi-pickups (purple dashed line) and check-ins (red dotted line) and make these additional observations:

(1) Unlike in the case of recent observations in Singapore, venue check-ins performed the best with capturing at least 50% of the events within 1.5 km of the event venue during the event start hour. We believe that this may be due to the popularity of the platform during the initial years of its existence which resulted in more dense check-in behavior.
(2) Although the two taxi signals exhibit low accuracy, we see that the taxi pickups show better accuracy closer to the end of the event compared to drop-offs, as anticipated.

4.5 Sensor Fusion

Here, we investigate the efficacy of fusing the disparate sensor sources in the context of event detection. Following the notations of the earlier analyses, we set the distance threshold, $R = 1.5km$, and the threshold for declaring an anomaly as $S = 0.8$ on the scaled intensity values between 0 and 1. In Figure 5, we plot the recall performance on the $y-$axis whilst the $x-$axis represents the mixture

[5]http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
[6]http://www1.nyc.gov/site/planning/data-maps/open-data/districts-download-metadata.page
[7]www.nycinsiderguide.com

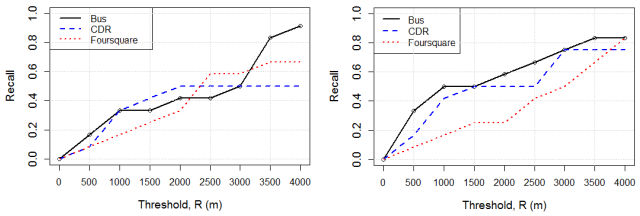

Figure 3: The trade-off between event recall and localization accuracy during (1) the event start hour (*left*) and (2) the hour prior to that (*right*).

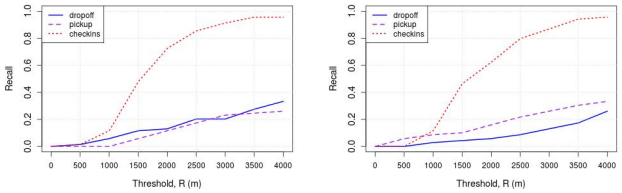

Figure 4: The trade-off between event recall and localization accuracy during (1) the event start hour (*left*) and (2) three hours later (*right*) on the NYC dataset.

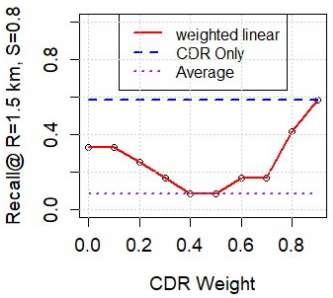

Figure 5: Recall performance with sensor fusion.

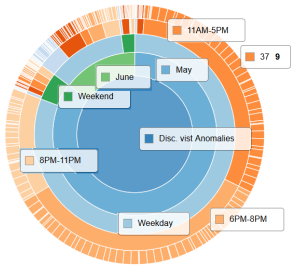

Figure 6: Sunburst view of anomalies

weight for CDR assuming a weighted linear combination. Here, we fix the weight of Foursquare to 0.1 due to its low performance. The dashed blue line represents the performance line using CDR alone and the dotted purple line represents the performance of the arithmetic mean of the three sources. We make the following interesting observations:

(1) With the current set of events, CDR shows the best performance. Interestingly, the average across the signals performs the worst.
(2) With increasing mixture of the CDR, we see that the performance drops till equal contribution, and then increases monotonically. This suggests that the bus data and CDR data, in fact, detect different sets of events highlighting the need for smarter fusion.

5 IMPACT ON PROBLEM

We list the potential impact and use cases of such a event detection and characterization web portal.

Figure 7: *Left* -Estimated and actual event venues of the Britney Spears Concert. Black - Actual, Orange - CDR, Green - Bus, Blue - Foursquare. *Right*-Word cloud during the Britney Spears Concert. Hashtags like "BritneySpears" were present.

General public: users will be able to query for current events related to their interests (e.g., condo launch) or by locality (e.g., events in the YCK neighborhood). This allows for serendipitous exploration in the event landscape, as an alternative to advertised and/or ticketed events on channels such as SISTIC[8]. Similarly, being able to visualize the impact of a road incident allows commuters to change their routes such that they avoid congested and possibly affected streets.

Event planners and Land Use Authorities: the generative nature of the event model enables future event organizers and resource planners to visualize and understand the impact of an event in the planning. This may help reveal unseen bottlenecks or unexpected anomalies that the event may cause.

6 WEB APPLICATION

The web application was prototyped using RShiny[8] and entails the following views.

Event Landscape View: This view,shown in Fig 1, provides the user a quick snapshot of the events landscape all over Singapore in the form of a series of images in time-lapse.

Summary of anomalies: This interactive view is shown in Fig 6, and it provides a summary of the anomalies observed in the CDR data. The first three concentric rings represents time units such as month, weekday/weekend and time-bins. The last ring represents the subzone and the count of anomalies. The size of a slice in a ring is proportional to the corresponding number of anomalies.

Data Layer View: This view provides the user with an interface to add/remove data sources used to perform the analysis.

Past Events View: This view, provides the user with overview of events that were detected in past. In Figure 7, we share a screenshot from the EventXplore dashboard which shows the localization accuracy of each source compared to the actual event venue for one of the largest events from our list – i.e., the Britney Spears concert.

7 COMPARISON TO RELATED WORK

Detecting anomalies in urban mobility patterns from physical sensors such as GPS traces and traffic cameras [1, 7], and CDR [10, 11] is a well-studied topic in the context of optimizing the traffic related infrastructure. CDR data have also been used to detect unusual urban events (e.g., elections, emergency events, etc.) [2, 5]. However, these studies are largely unimodal in nature. A number of works that exploit multimodality have emerged; however, these focus predominantly on anomalies related to traffic (e.g., accidents). For example, the combination of sparsely available GPS data and Tweets were used in [9] to observe the congestion along road segments. The authors in [4] make the case for identifying root causes for sensor anomalies using social media data. Recently, multimodal sensing approaches have been attempted for urban event detection. A two-step modeling process was used in [6] to predict irregularities (e.g., large scale events) from multimodal data. However, this relies on App usage data that is only predictive of planned/anticipated events of large scale.

In this work, we focus on the problem of detecting and gaining insights into urban events of varying scale (small, large and medium) and we investigate the utility of the various sensors in providing fine grained and clear insights into such urban events. Apart from detecting and localizing the events in space and time, we also investigate the possibility of providing a semantic annotation of the detected events.

Concluding Remarks. In this work, we investigated the ability of multimodal urban data sources in detecting events both spatiall and temporally. As observed, we intend to develop and incorporate smarter fusion algorithms that accounts for spatial and temporal estimation biases for more accurate detection and annotation. The video illustrating the web application we developed is available from: http://ares.smu.edu.sg/~kasthurij/videos/cikm-video.mp4.

[8]https://shiny.rstudio.com/

REFERENCES

[1] Sanjay Chawla, Yu Zheng, and Jiafeng Hu. 2012. Inferring the root cause in road traffic anomalies. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 141–150.

[2] Yuxiao Dong, Fabio Pinelli, Yiannis Gkoufas, Zubair Nabi, Francesco Calabrese, and Nitesh V Chawla. 2015. Inferring unusual crowd events from mobile phone call detail records. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 474–492.

[3] Elastic.co. 2017. RESTful, Distributed Search and Analytics. https://www.elastic.co/products/elasticsearch. (2017). Accessed: August 21, 2017.

[4] Prasanna Giridhar, Md Tanvir Amin, Tarek Abdelzaher, Dong Wang, Lance Kaplan, Jemin George, and Raghu Ganti. 2016. ClariSense+: An enhanced traffic anomaly explanation service using social network feeds. *Pervasive and Mobile Computing* 33 (2016), 140–155.

[5] Didem Gundogdu, Ozlem D Incel, Albert A Salah, and Bruno Lepri. 2016. Countrywide arrhythmia: emergency event detection using mobile phone data. *EPJ Data Science* 5, 1 (2016), 25.

[6] Tatsuya Konishi, Mikiya Maruyama, Kota Tsubouchi, and Masamichi Shimosaka. 2016. CityProphet: city-scale irregularity prediction using transit app logs. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 752–757.

[7] Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan, and Xie Xing. 2011. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1010–1018.

[8] SISTIC.COM. 2013. Singapore Event and Tickets Online. (March 2013). Retrieved August 21, 2017 from https://www.sistic.com.sg/

[9] Senzhang Wang, Lifang He, Leon Stenneth, S Yu Philip, Zhoujun Li, and Zhiqiu Huang. 2016. Estimating urban traffic congestions with multi-sourced data. In *Mobile Data Management (MDM), 2016 17th IEEE International Conference on*, Vol. 1. IEEE, 82–91.

[10] Peter Widhalm, Yingxiang Yang, Michael Ulm, Shounak Athavale, and Marta C González. 2015. Discovering urban activity patterns in cell phone data. *Transportation* 42, 4 (2015), 597–623.

[11] Mogeng Yin, Madeleine Sheehan, Sidney Feygin, Jean-François Paiement, and Alexei Pozdnoukhov. 2017. A generative model of urban activities from cellular Data. *IEEE Transactions on Intelligent Transportation Systems* (2017).