Profiling Pedestrian Activity Patterns in a Dynamic Urban Environment

Minh Tuan Doan, Sutharshan Rajasegarar, Christopher Leckie

National ICT Australia, Department of Computing and Information Systems

The University of Melbourne, Australia

Email: {mdoan@student., sraja@, caleckie@}unimelb.edu.au

ABSTRACT

In this paper we model the normal behaviour of pedestrians over the central business district of the City of Melbourne over a long period, and then use this model to detect the anomalous distributions. We develop an approach based on frequent itemset mining as the underlying technique and construct a model in terms of correlations of pedestrian distributions across multiple locations in the city. The algorithm is demonstrated to construct a model that is reliable and representative of the dataset. The modelling and anomaly detection are then both empirically evaluated at a high level and cross referenced with known events. The results show that anomalies can be detected with high reliability using our approach.

1. INTRODUCTION

The emergent of Internet of Things (IoT), which is a major evolution of the current Internet into a network of interconnected objects, has resulted in many sensors being deployed to monitor various parameters of interest in the city [3, 12, 21, 14]. Examples of such deployments include SmartSantander [5], the City of San Francisco [4] and the City of Melbourne, Australia [1]. Pedestrian activities play an important role in the dynamics of cities. The involvement of pedestrians in almost every operation of the city, such as infrastructure, transportation and traffic has made understanding their distribution an essential task for architects, city planners or event schedulers. The pedestrian movement data collected using sensors in the city enable the pattern of movements of pedestrians to be studied in a systematic way. However, due to the volume and complex structure of the data, there has been limited progress in utilizing sensor data to model pedestrian behaviour and detect any interesting events.

In this study we aim to model the activities of pedestrians over the central business district of the City of Melbourne, Australia. By mining the correlations of pedestrian distributions across multiple key locations of the city, we define a profile of normal pedestrian distributions, and then use this model to detect any anomalous pedestrian distributions. This knowledge is valuable to city planners, business managers and many other audiences to better understand urban dynamics, which can facilitate more efficient ways of designing, planning and scheduling of events in large cities.

We employ a frequent itemset mining [7] based approach to reveal the correlation of pedestrian loads at various locations in the city during a given time of the day. By constructing the set of frequent itemsets that represent the majority of the observations in the dataset, we build a reliable model that describes the normal state of pedestrian behaviours. Consequently, we use the model to detect any anomalous activities of pedestrians. There are two main research problems that need to be addressed in this study. First, the data is multidimensional and the raw values of pedestrian counts in each dimension (location) might carry a different meaning in terms of the distribution when compared with the other dimensions, e.g., the count of pedestrians collected at a busy train station might have a different significance than if the same value is observed at a vacant park. Second, modelling pedestrian data is not purely a task of frequent itemset mining. In addition to producing reliable frequent itemsets with high support, the algorithm needs to ensure that the set of frequent itemsets are representative of not only the observations in the dataset but also of the locations being investigated. Only by fulfilling these two requirements will the model be likely to be meaningful and the anomaly detection accurate.

The data used in this study is pedestrian counting data from the City of Melbourne, Australia [1], where sensors were deployed to capture pedestrian counts over 28 locations around the centre of the city. The counts are collected continuously at intervals of 1 hour and provide a high level overview of the pedestrian distribution in the Melbourne CBD on an hourly basis. Our key contribution is that we transform the raw and discrete data of pedestrian counts into a meaningful model that describes the patterns of pedestrian activities. In addition we used the model to detect anomalous events and validated them with known events.

The rest of the paper is organized as follows. Section 2 describes the related work in profiling pedestrian movements.

UrbComp'15, August 10, 2015, Sydney, Australia.

Section 3 defines the analysis problem and presents the underlying challenges. Section 4 presents our algorithm in modelling the pedestrian distribution and detecting anomalies. Section 5 describes the experiments and results. Section 6 gives an overall conclusion.

2. RELATED WORK

Most studies on pedestrian mobility patterns have used GPS or ubiquitous devices to track the locations of pedestrians and analyzed them on a microscopic level. These studies focused mainly on user-centric activities such as individual trajectories or activities [22][18].

There are however, relatively few studies on the crowd distribution and pedestrian mobility of a large metropolitan region as a whole. Reades et al. use Erlang [20] cellular data to analyse the distribution of phone users across multiple locations of the City of Rome [20], and to construct the crowd distribution of these locations. The data provided by the telecommunication company was collected over 90 days at intervals of 15 minutes, and it provides a spatio-temporal overview of the phone usage at different locations in the city. The study managed to differentiate between usage profiles during weekdays and weekends, to explain some routine activities of the city such as sport events on weekends, leisure activities in late evening or transition states of the city on Friday afternoons. The authors also performed the clustering of geographical locations based on their corresponding distributions of phone users at different times of the day and showed that it can separate the center vs. the outskirts of the city, and highlighted the hot spots of the city.

Many studies have used geo-tagged data from location based social networks such as Facebook, Twitter or Flickr to model crowd distributions [13][17]. For example, Ferrari et al. [11] model the crowd footprint of Manhattan using geo-tagged Twitter data. The authors associated the crowd footprint to the words used in messages and used Latent Dirichlet Allocation [9] to reveal the trends and routine behaviour in the area for each day of the week. However, in this study the number of pedestrian load profiles need to be predicted in advance and the result is quite sensitive to this input. Moreover, the profiles were assumed to be constant throughout the one year span of the collected data, which may result in low accuracy as presented in the paper. For example, by validating all Mondays with a typical Monday profile, the accuracy yields less than 40%.

Kaltenbrunner et al. [15] studied the usage of public bike sharing services across multiple locations in the City of Barcelona. By investigating the number of bikes rented and available at the bike sharing stations, the study inferred the cycling activities, analyzed the mobility patterns and modelled the cycling activities of Barcelona's population. The study used 7 weeks worth of data which had been directly retrieved from the service provider. When analyzing the data at each location separately, the authors distinguished two different profiles for weekdays and weekends and justified these patterns with routines in the city, such as working hours, lunch breaks, etc. The study also analyzed the data on a macroscopic level and revealed the global mobility pattern of Barcelona's cyclists. The result is expressed as a geographic heat map which spatio-temporally present the increase and decrease of cycling activities over the entire city. This result, however, can only be interpreted visually and does not really facilitate further analysis in a systematic and programmatic way. In addition, the metrics used in this study to deduce the profiles are fairly simple, e.g., average, standard deviation, local maximum and local maximum. We believe more complex analysis on this type of data can reveal more detailed patterns.

The studies mentioned above use a short period of collected data where it can safely be assumed that the system stays in one state. Moreover, although they study the patterns across multiple locations, they do not perform analysis at all the locations simultaneously, which we believe will provide valuable information about the correlations between different locations. In this paper, we focus on analyzing the high dimensional data that spans over a long period to reveal these correlations.

3. PROBLEM STATEMENT & CHALLENGES

3.1 Problem Statement

Let O denote the set of N observations: $O = \{o_i : i = 1..N\}$ where $o_i \in \mathbb{R}^d$ is a vector of d dimensions corresponding to the pedestrian count values at d different sensor locations. Furthermore, o_{pl} represents the observation at time t_p at location l, where $l = 1..d$.

First, we aim to model the normal behaviour of O as the set of correlations of pedestrian distributions at different locations during the same period of the day. The model F_p is defined as the *frequent itemsets* of the input data subset O_p containing the pedestrians distributions at time t_p of the day:

$F_p = \{X_{pi} : i = 1..K\}$, X_{pi} is a frequent itemset of O_p.

Second, we aim to find the set A containing all the anomalies that are not covered by any frequent itemsets.

By evaluating the set of models F_p at different time t_p of the day, we seek to understand the correlations of pedestrian loads at different locations of the city and use that knowledge to systematically validate and explain the anomalies found in set A.

3.2 Challenges

In order to describe the distribution of pedestrians across the city, we consider observations at multiple locations in the city, resulting in high dimensional input data. Our initial attempts in detecting anomalies using clustering algorithms on this high dimensional dataset has been unsuccessful. According to Kriegel et al. [16], clustering of high dimensional data poses multiple challenges for clustering algorithms because traditional similarity measures become sensitive to noise. Challenges also arise from local feature correlations in high dimensional data, which entails that clusters might exist in different subspaces but not in the global feature space being considered. For example, the pedestrian counts at Flinders Street Station, Princes Bridge and the Arts Center might have strong correlations as they are located near each other. However, the values collected by more distant sensors at the Queen Victoria Market or China Town, might

not be relevant, and might possibly hinder the formation of clusters. Besides the high dimensionality, the data used in this study is not evenly distributed, with a large proportion of the observations being skewed towards the lower values, which makes clustering even more difficult.

Another challenge in modeling the data comes from the long time scale of the data being collected, which might cause the model of normal behaviour to vary within different time frames as pedestrian behaviours may change intermittently throughout the year.

To overcome these challenges in order to detect anomalies effectively, we aim to develop an algorithm that (1) can handle high dimensional data with attributes that share different degrees of correlation and (2) can cope with a system that switches between different states over time, e.g., morning peak, lunch time and afternoon peak.

Moreover, modelling the pedestrian activities are more complicated than the task of pure frequent itemset mining. One of the major challenges is that the output frequent itemsets that form the model need to be both reliable, meaningful and representative of the dataset. Consequently, the following constraints were imposed:

Constraint 1: The support $\sigma(X_i)$ of each frequent itemset X_i needs to be high to ensure reliable sets.

Constraint 2: The combination of frequent itemsets in F_p needs to cover a significant proportion of observations of the input dataset. We assume that normal behaviour is the default state that describes the majority of observations, hence a high coverage ensures that the model does not miss any interesting patterns in the data.

Constraint 3: The union of all items (locations) of the frequent sets need to cover a sufficient proportion of locations being investigated. This requirement prevents the constraints (1) and (2) from being overlooked by small frequent itemsets. For example, it is observed from our experiments that O_p, for different values of p, has many small frequent 2-itemsets with very high support (near 100%); if a high `min_sup` (minimum support threshold) is selected, constraints (1) and (2) are easily satisfied by these 2-itemsets. However, in these cases, the model is dominated by itemsets of only a few locations and many other locations were left out, which leads to many anomalies at these locations failing to be detected.

4. METHODOLOGY

This section describes our method to model the normal pedestrian distribution and to detect anomalous behaviours. The algorithm comprises two phases - data normalisation and modelling normal behaviour.

4.1 Data normalization

The count values collected by the sensors offer few insights on the correlations between pedestrian distributions at different locations. Moreover, as mentioned, since the system can change between different states intermittently, comparing the values of the same sensor but at different times can also lead to a misleading estimate of the true distribution. For example, a count of 1500 pedestrians is unusually high for the Queen Victoria Market. However, at Flinders Street Station, it is considered low during most times of the year. Data normalization aims to better reveal the relationships of the count values at different locations in different time frames.

The stream of data is divided into windows of size w. In each window, the maximum count at each location is chosen as the standard measure (100%) and all the remaining values are classified into buckets of `LOW`, `MEDIUM`, `HIGH` in proportion to their corresponding measure using the scale below:

$[0\%, 30\%) \rightarrow$ `LOW`

$[30\%, 70\%) \rightarrow$ `MEDIUM`

$[70\%, 100\%] \rightarrow$ `HIGH`

By dividing the data into small windows before classifying the pedestrian counts into buckets, it is observed that the classification reflects the correlations more accurately because it is assessed locally in a short time frame and is not biased by factors that change over time, such as weather or holiday periods. This step of normalization also makes the algorithm more robust when the system switches between states.

As an illustration, the data normalization process is applied on the data collected by the sensors at Flinders Street Station and Southern Cross Station during the morning peak between 1^{st} Jun and 31^{st} Dec 2014. The process is shown in Table 1. This example also demonstrates the importance of the normalization to express the true nature of the pedestrian distribution at each station so that their correlations can be analyzed more accurately. As shown in Table 1, the counts of 769 pedestrians at Flinders Street Station and 773 at Southern Cross Station are very similar, however they represent two different distributions at two locations: LOW at the former and $MEDIUM$ at the latter location.

The labelled data is then binarized [10] before we can apply frequent itemset mining in the next phase to model the normal behaviour. In this step, each location L is then assigned with three attributes L_{LOW}, L_{MED} and L_{HIGH} which express the presence or absence of that type of distribution at the location.

4.2 Normal Behaviour Modelling and Anomaly Detection

The Apriori algorithm [8] is used as the underlying technique to mine frequent itemsets from the input. The main challenges of the modelling process lie in selecting an appropriate support threshold that balances the coverage of the observations, the coverage of locations and the reliability of the frequent itemsets as mentioned in the problem statement. A support threshold that is too low increases the chance that multiple locations will be included into the frequent sets and more observations will be covered, but may result in frequent itemsets that are not reliable. On the other hand, a support threshold that is too high guarantees that the frequent sets are strong, but at the same time might exclude more locations from the frequent sets. The algorithm used in this

Count values	142	442	1089	2338	2280	1423	1068	219	769	849	1625	...	1829	2040	2914	2446
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓		↓	↓	↓	↓
Percentage to max value	4.8	15.1	37.3	80.2	78.2	48.8	36.6	7.5	26.3	29.1	55.7	...	62.7	70.0	100	83.9
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓		↓	↓	↓	↓
Bucket	L	L	M	H	H	M	M	L	L	L	M	...	H	H	H	H

(a) Flinders Street Station

Count values	7	3	117	1419	1987	2404	2235	773	349	1115	1802	...	377	337	479	441
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓		↓	↓	↓	↓
Percentage to max value	0.3	0.1	4.9	59.0	82.7	100	92.9	32.2	14.5	46.3	74.9	...	15.7	14.0	19.9	18.3
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓		↓	↓	↓	↓
Bucket	L	L	L	M	H	H	H	M	L	M	H	...	L	L	L	L

(b) Southern Cross Station

Table 1: Example of data normalization of pedestrian counts at Flinders Street and Southern Cross Station for the data window 15th Dec - 31st Dec 2014. The maximum value of the window is highlighted in grey.

study to achieve an appropriate mix of frequent itemsets is described in Algorithm 1.

The algorithm attempts to find the highest support threshold for which the frequent itemsets achieve a good coverage of the locations as well as represent a majority of the observations. The algorithm is initialized with the highest support threshold `min_sup`. Then in each iteration, `min_sup` is decreased and the corresponding frequent itemsets are generated and tested to see whether they provide a good model. The algorithm stops when the first set of frequent itemsets that satisfies all the requirements above is found, or when `min_sup` reaches the lower bound threshold. In this study, we set the threshold of records to be covered $rCoverage = 80\%$, the threshold of locations to be covered $lCoverage = 70\%$ and the support decrement step to 5%.

Any records that do not match any of the frequent itemsets are considered as anomalies.

Next, we present an illustrative example of the application

Algorithm 1: Modelling normal behaviour

Input : O_p^*: set of N normalized observations at time t_p of the day
$rCoverage$: threshold of records to be covered
$lCoverage$: threshold of locations to be covered
$sThreshold$: threshold of minimum support

Output: $F_p = \{X_i : i = 1..K\}$: set of all frequent itemsets that satisfy the constraints

```
min_sup := 1;
while (notfound and min_sup > sThreshold) do
    F = apriori(O_p^*, min_sup)
    //remove the 2-itemsets from F
    F :=remove2ItemSet(F);
    //test the coverage of the model
    cl := findLocationCoverage(F, lCoverage);
    cr := findRecordCoverage(F, rCoverage);
    If (cl > lCoverage and cr > rCoverage);
    found := true;
    else min_sup := min_sup − 0.05 ;
end
return F
```

Iteration	Outputs	Constraint checks
1	$\emptyset$	$rC = 0\%$ $lC = 0\%$
2 3 4 5 6	$\{PB\_LOW, SC\_LOW\}$ (discarded)	$rC = 98\%$ $lC = 50\%$
7	$\{FL\_MED, PB\_LOW, SC\_LOW\}$	$rC = 73\%$ $lC = 75\%$
8	$\{FL\_MED, PB\_LOW, SC\_LOW\}$ $\{PB\_LOW, SC\_LOW, AC\_MED\}$	$rC = 88\%$ $lC = 100\%$

Table 2: Illustration of the iterations of Algorithm 1. rC, lC are the coverage level of the frequent itemsets on the records and locations respectively.

of the algorithm on a subset of the data. The input is the 4-dimensional data collected by the sensors at Flinders Street Station (FL), Southern Cross Station (SC), Princes Bridge (PB) and the Arts Center (AC) during morning peaks (7am-9am) from 1^{st} Jun to 31^{st} Dec 2014.

The data was first normalized and used as the input of Algorithm 1. The frequent itemsets that are output as well as their corresponding coverage on the observations and on the locations in each iteration are shown in Table 2.

Initially, `min_sup` is set to 100% and no frequent itemsets can be found as expected. The minimum support threshold is then decremented at a step of 5% in each iteration. In the next six iterations, where `min_sup` is decreased from 100% to 75%, only one frequent itemset can be mined from the data, but it contains only two items and is discarded anyway. In this study we decided to ignore all the 2-itemsets in the output because they create bias when judging how many observations are covered by the frequent itemsets. It was observed from our experiments that in most of the cases, the 2-itemsets have very high support, i.e., they cover a large proportion of the data, which makes **Constraint 2** less meaningful. Including these small sets might lead to a poor model that contains only 2-itemsets. In this case, all three constraints are satisfied but the model only barely

contains correlations in terms of pairs of locations, rather than the correlations of pedestrian distributions at multiple location as a whole.

In the 7th iteration, with `min_sup` decreased to 70%, the output 3-itemset satisfies **Constraint 3** but covers fewer observations than required. In the 8th iteration, the Apriori algorithm finds two 3-itemsets that, in combination, cover 88% of all the observations. Moreover, all locations are covered by this output. All three constraints are met and the algorithm terminates. The itemsets found in the 8th iteration form the model that describes the normal correlations of pedestrian distributions at Flinders Street Station, Southern Cross Station, Princes Bridge and the Arts Center during the morning peak.

5. EVALUATION AND RESULTS

In this section, we empirically evaluate our approach for anomaly detection based on frequent itemset mining for pedestrian modeling.

5.1 Evaluation

We perform an experiment on the pedestrian counts from the sensors deployed at 10 different locations around the City of Melbourne. As illustrated in Figure 1, these monitoring points are located at the major train stations, city hubs or shopping malls and are representative of the flows of pedestrians inward, outward and within the CBD area. The time period of the monitored data is from 1st Jun to 31st Dec 2014.

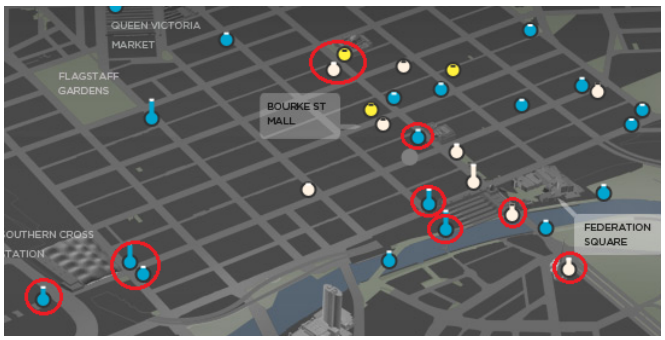


Figure 1: Deployment map of sensors (sensors circled in red are used in this study). Figure reproduced from the website of Pedestrian Counting System of the City of Melbourne [1].

In this study we focus on 2 periods of the day: morning peak (7am-9am) and midnight (10pm-midnight). The subsets of observations $O_{morning}$ and $O_{midnight}$ were each used as input for the modelling and anomaly detection. We conducted two phases of verification on the results of anomalies:

1. *High level verification:* In this phase we empirically compare the anomalies against the model of normal behaviours. The differences between the two sets of events can provide a high level overview of the effectiveness of the algorithm, since it is expected that anomalies will strongly show different attributes compared to the model of normal behaviours.
2. *Concrete verification:* A sample of the reported anomalies is selected and we manually validate whether they are actually anomalies. Due to the lack of ground truth labels, this step of verification requires us to manually cross reference each anomaly of the sample with known events.

5.2 Result and Discussion

5.2.1 High level verification

We found 21 anomalies spread over 10 different days during the morning peaks and 91 anomalies spread over 60 days during midnight periods. The statistics of the `LOW`, `MED` and `HIGH` distributions of these anomalies were compared with those defined as the normal behaviours. The comparison at Flinders Street Station and Southern Cross Station during morning peaks is shown in Table 3.

		Flinders Station			Southern Cross		
		L	M	H	L	M	H
Normal	7am	11	142	0	153	0	0
	8am	7	86	60	83	70	0
	9am	38	115	0	18	135	0
Anomalies	7am	9	0	0	8	2	0
	8am	7	0	0	6	1	0
	9am	5	0	0	4	0	0

Table 3: Comparison of pedestrian distribution at Flinders Street and Southern Cross Stations.

The differences between the two sets of normal and anomalous records are clear. Flinders Street Station usually observes `MED` distribution at 7am, then increases towards `MED-HIGH` at 8am and then decreases back to `MED` at 9am. However, the anomalous records show totally different trends where the distributions always remain `LOW`. The same clear contrast can also be observed from the values of Southern Cross Station.

The contrast between normal and anomalous records at all 10 locations are visualized in Figure 2. We display the distribution of pedestrians as an intensity image. For time t_p of the day, each location is represented by a block of gray tone. The intensity of the gray tone is proportional to the average level of the pedestrian distribution at time t_p in the dataset, which is defined as:

$$P_p = \frac{\alpha_1|LOW| + \alpha_2|MEDIUM| + \alpha_3|HIGH|}{|LOW| + |MEDIUM| + |HIGH|}$$

where $|LOW|$, $|MEDIUM|$, $|HIGH|$ are respectively the numbers of `LOW`, `MEDIUM` and `HIGH` distributions and α_1, α_2, α_3 are the corresponding weights to distinguish the contributions of each type of distribution to the intensity image.

The differences can be observed at almost every location for each time slot. The contrasts between the normal (N-7am, N-8am and N-9am) and anomalous records (A-7am, A-8am and A-9am) shown in Figure 2 indicate that the different behaviours are significant. For example, during the anomalous events at Town Hall West, the number of pedestrians is higher than usual at 7am, it then decreases to slightly lower than usual at 8am, and to a much lower level at 9am.

5.2.2 Concrete verification

In this section we select a sample of the anomalies detected by the algorithm and attempt to validate them against known

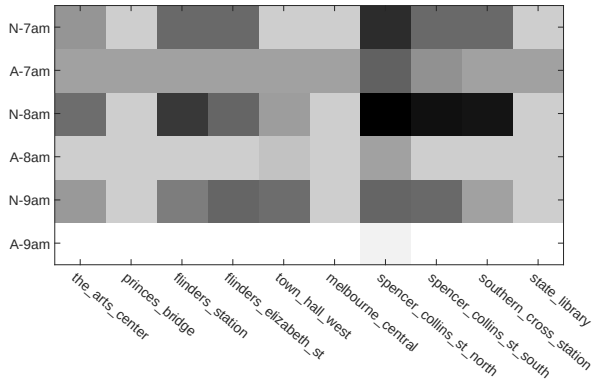


Figure 2: Contrast between normal behaviour and anomalous events at 10 locations during morning peaks. Blocks with darker grey represent higher distributions of pedestrian at the given location and vice versa.

events. An anomaly can be validated by being matched to a special event such as a public holiday, or an anomaly in the weather. Three hours were used to define the morning peak and the midnight period in this study, and they also contribute to the ranking of anomalies. If anomalies are observed throughout all the hours during the period, they are considered to be more robust and receive a higher ranking. The anomalies detected during the morning peak period are listed and explained in Table 4.

Date	Count of anomalies	Justification
2014.06.09	3	Queen's Birthday
2014.11.04	3	Melbourne Cup
2014.12.25	3	Christmas
2014.12.31	3	New Year's Eve
2014.12.26	2	Boxing Day
2014.12.29	2	Holiday period
2014.12.30	2	Holiday period
2014.09.22	1	Holiday period
2014.12.23	1	Holiday period
2014.12.24	1	Christmas Eve

Table 4: Anomalous records detected during morning peak periods.

Most of the anomalies can be matched to major events during the year when the pedestrian distribution is expected to be different from the norm. During public holidays, less people are rushing to the city for work at early hours in the morning, which explains the unusually lower distribution at most of the locations as shown in Figure 2. Although the dates 2014.12.22, 2014.12.23, 2014.12.29, 2014.12.30 do not match to any specific events, they are the Mondays and Tuesdays of the holiday season, right before Christmas and New Year. As many business close over this period, it is arguable that fewer people come to the city in the morning on these dates.

During the midnight period, 91 anomalies spread over 60 days were detected. Using the event listing website [6, 2], we matched many of these anomalies to sports events or festivals that happened at night time. A summary of the results are shown in Table 5.

Date	Justification	Date	Justification
5th Dec	Food Festival	31st Dec	New Year's Eve
24th Dec	Christmas' Eve	27th Nov	Football match
21st Nov	Cricket match	20th Nov	Equestrian event
12th Nov	Football match	7th Nov	Cricket match
31st Oct	Halloween	10th Oct	Melboure Festival
Summary : 56 anomalies fall on Fridays 18 anomalies can be matched with an event 17 anomalies do not match with any event			

Table 5: Anomalous records detected during midnight periods.

It is observed that most of these anomalies happen on Friday night. It it natural that more pedestrians are expected to show up in the city area on Friday nights, causing anomalous distributions. Due to the lack of ground truth labels, Friday nights can be used as good indicators to measure the effectiveness of the algorithm. As 56 anomalies spread over 29 Fridays were found, out of 30 Fridays from 1st June to 31st Dec 2014, this can partially confirm the accuracy and high recall rate of the algorithm. There were 17 anomalies that we could not match to any major events but they are not necessarily false positives. The unusual distribution of pedestrians could have been caused by the weather, transportation disruptions or any unofficial events, for which it is challenging to find references.

5.3 Time Complexity

In this section we discuss the time complexity of the algorithm when it is applied on N observations of pedestrian counts collected by sensors at l locations.

The data normalization traverses all the observations twice, to find the maximum values of each window then to assign each pedestrian count values into buckets of `LOW`, `MEDIUM` and `HIGH`, hence the time complexity of the data normalization step is $O(N * l)$.

The testing of the coverage of the frequent itemsets on the dataset in each iteration of Algorithm 1 is also linear in the number of observations:

$O(N * f)$, where f is the number of frequent itemsets found and $f \ll N$.

We observed from our experiments that it is the mining process for frequent itemsets at each iteration of Algorithm 1 that takes the majority time of the modelling process. According to Pang-Ning [19] et al., multiple factors can affect the complexity of the Apriori algorithm, including the support threshold, dimensionality of the dataset and number of observations. Since our algorithm executes the Apriori algorithm several times until all the requirements are satisfied, the number of iterations as well as the values of the trial supports can only be known in retrospect. It is hence challenging to analytically formulate the time complexity of the algorithm.

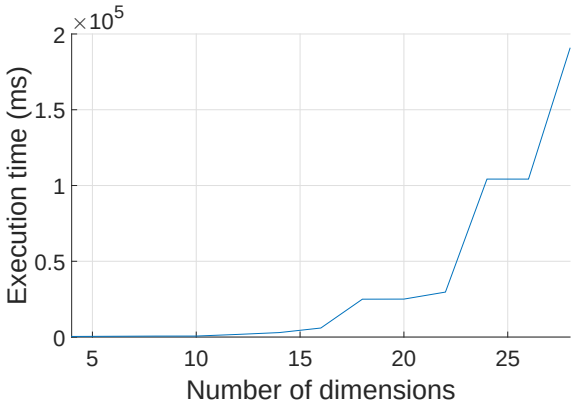
Consequently, we have empirically analyzed the effect of the

number of observations N and the dimensionality l on the time complexity. We tested the algorithm on four subsets of data: 3 months, 6 months, 9 months and 12 months worth of data, and observed that there is little variation in the observed execution times as the number of observations increases. We then tested the algorithm on data with varying numbers of sensor locations (dimensions), ranging from 4 to 28 dimensions. For each dimension, the algorithm was applied on all four subsets of the data mentioned above. The average execution time for each number of dimensions is presented in Table 6 and plotted in Figure 3.

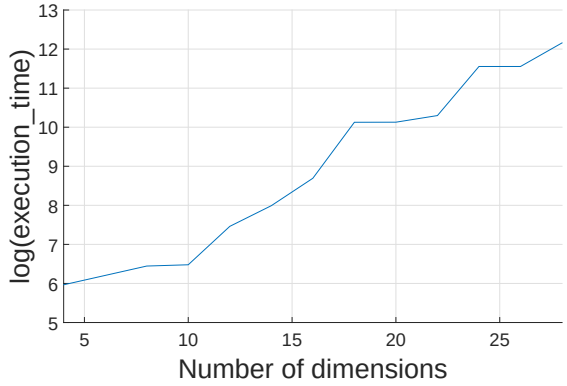
Dimensionality (number of locations)	Execution time (ms)
4	390
8	630
10	651
12	1740
14	2954
16	5960
18	24987
20	25044
22	29675
24	104241
26	104227
28	190620

Table 6: Execution time of the algorithm with different numbers of dimensions.

Figure 3 shows that the execution time grows exponentially with the number of dimensions, i.e. the number of locations being investigated.



(a) time in linear scale.



(b) time in log scale.

Figure 3: Execution time of the algorithm with different dimensionality.

6. CONCLUSION

In this paper we used frequent itemset mining as the underlying technique to form a model of pedestrian activities as the correlations of the distributions across multiple locations. The model was then used to extract anomalous events from the dataset, which were subsequently validated against known events. The empirical results indicate that the model is reliable and can be used to detect anomalies with high effectiveness.

7. ACKNOWLEDGMENT

The authors would like to thank the support from National ICT Australia (NICTA).

8. REFERENCES

[1] City Of Melbourne : Pedestrian Counting System. `http://www.pedestrian.melbourne.vic.gov.au/`.

[2] Events Calendar for 2014 | Victorian Government. `http://www.vic.gov.au/calendar/2014/01.html`.

[3] Internet of Things (IoT) - ISSNIP. `http://issnip.unimelb.edu.au/research_program/Internet_of_Things`.

[4] San Francisco parking data. `http://sfpark.org`.

[5] SmartSantander. `http://www.smartsantander.eu`.

[6] What's on in Melbourne December 2014. `http://www.victoria.aussietrueblue.com/what-to-do-in-Melbourne-December-2014.php`.

[7] C. Borgelt. Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):437–456, 2012.

[8] C. Borgelt and R. Kruse. Induction of association rules: Apriori implementation. In *Compstat*, pages 395–400. Springer, 2002.

[9] D. M. Blei et al. Latent Dirichlet allocation. *the Jnl. of Machine Learning Research*, 3:993–1022, 2003.

[10] T. Elomaa, J. Hollmén, and H. Mannila. *Discovery Science: 14th International Conference, DS 2011, Espoo, Finland, October 5-7, Proceedings*, volume 6926. Springer Science & Business Media, 2011.

[11] L. Ferrari et al. Extracting urban patterns from location-based social networks. In *ACM SIGSPATIAL Intl. Workshop on Location-Based Social Networks*, LBSN '11, pages 9–16, 2011.

[12] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami. Internet of things (iot): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7):1645–1660, 2013.

[13] L. Hollenstein and R. Purves. Exploring place through user-generated content: Using Flickr tags to describe city cores. *Jnl. of Spatial Information Science*, (1):21–48, 2015.

[14] J. Jin, J. Gubbi, S. Marusic, and M. Palaniswami. An information framework for creating a smart city through internet of things. *Internet of Things Journal, IEEE*, 1(2):112–121, 2014.

[15] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4):455–466, 2010.

[16] H.-P. Kriegel, P. Kröger, and A. Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):1, 2009.

[17] M. Mamei et al. Automatic analysis of geotagged photos for intelligent tourist services. In *Intelligent Environments*, pages 146–151, 2010.

[18] A. Millonig and G. Gartner. Shadowing, tracking, interviewing: How to explore human spatio-temporal behaviour patterns. In *BMI*, pages 1–14, 2008.

[19] T. Pang-Ning, M. Steinbach, V. Kumar, et al. Introduction to data mining. In *Introduction to data mining*, page 74, 2006.

[20] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti. Cellular census: Explorations in urban data collection. *Pervasive Computing, IEEE*, 6(3):30–38, 2007.

[21] A. Shilton, S. Rajasegarar, C. Leckie, and M. Palaniswami. Dp1svm: A dynamic planar one-class support vector machine for internet of things environment. In *2015 International Conference on Recent Advances in Internet of Things (RIoT)*, pages 1–6. IEEE, 2015.

[22] S. van der Spek. Spatial metro–tracking pedestrians in historic city centres. *Research in Urbanism Series*, 1(1):77–97, 2008.