Utilising Location Based Social Media in Travel Survey Methods: bringing Twitter data into the play

Alireza Abbasi
School of Engineering and Information Technology
UNSW Australia
a.abbasi@unsw.edu.au

Taha Hossein Rashidi
School of Civil and Environmental Engineering
UNSW Australia
rashidi@unsw.edu.au

Mojtaba Maghrebi
School of Civil and Environmental Engineering
UNSW Australia
maghrebi@unsw.edu.au

S. Travis Waller
School of Civil and Environmental Engineering
UNSW Australia
s.waller@unsw.edu.au

ABSTRACT

A growing body of literature has been devoted to harnessing the crowdsourcing power of social media by extracting knowledge from the huge amounts of information available online. This paper discusses how social media data can be used indirectly and with minimal cost to extract travel attributes such as trip purpose and activity location. As a result, the capacity of Twitter data in complementing other sources of transport related data such as household travel surveys or traffic count data is examined. Further, a detailed discussion is provided on how short term travellers, such as tourists, can be identified using Twitter data and how their travel pattern can be analysed. Having appropriate information about tourists/visitors – such as the places they visit, their origin and the pattern of their movements at their destination – is of great importance to urban planners. The available profile information of users and self-reported geo-location data on Twitter are used to identify tourists visiting Sydney as well as also those Sydney residents who made a trip outside Sydney. The presented data and analysis enable us to understand and track tourists' movements in cities for better urban planning. The results of this paper open up avenues for travel demand modellers to explore the possibility of using big data (in this case Twitter data) to model short distance (day-to-day or activity based) and long distance (vacation) trips.

Categories and Subject Descriptors

H2.8 [**Database Applications**]: Data mining, Spatial databases and GIS
H3.5 [**Online Information Services**]: Web-based services

Keywords

Social media; Travel demand analysis; Topic modelling; Location based social media

LBSN'15, November 03, 2015, Bellevue, WA, USA

ACM 978-1-4503-3975-9/15/11…$15.00
DOI: http://dx.doi.org/10.1145/2830657.2830660

1. INTRODUCTION

The history of planning transport system infrastructures goes back to when the wheel was invented, followed by the construction of the first paved road in Sumer in 500 BC. At the same time, Darius I the Great began construction of an extensive road system for Persia, including the famous Royal Road which was one of the first highways. About the same time, Roman roads were constructed with advanced technologies; they were stone-paved, cambered for drainage and flanked by footpaths, bridleways and drainage ditches. The same road structure was later used by Great Britain in the 18th century to establish the first toll system which included 250 miles of road and 40 bridges. All of these early transport system planning and network design efforts inspired transport engineers of the 20th century to develop systematic procedures for policy appraisal and network design purposes. In the 1950s the first prototypes of the conventional four-step models were developed in Chicago and Detroit in the USA. Since then, many metropolitan areas have adopted a similar structure to evaluate the short, medium and long term consequences of different designs and policies.

The four-step modelling paradigm, which is a trip-based approach, led to the tour-based scheme in which individual level travel information is regarded for modelling purposes. Tour-based models were later evolved to activity-based models in which individual/household level data is used to model individual/household level travel attributes. Travel demand models are crucial to city planners and policy makers because through these models long term plans are assessed and selected. Further, finding solutions for the consistently dropping level of service of the transport network in a systematic and analytical manner needs travel demand models.

The evolution of travel demand modelling techniques brought about the need for high resolution databases in which socio-demographic and economic attributes of people are used to model their day-to-day travel behaviour. Such data sources encompass the travel diaries of a sample of people representing the population. Having access to such an individual level travel diary is crucial in developing several components of advanced behavioural modelling frameworks such as tour-based and activity-based. The most important travel attributes considered in these modelling frameworks are:

1. Trip purpose
2. Departure time
3. Mode of transport
4. Activity duration

5. Activity location
6. Travel route
7. Party composition
8. Traffic condition

Traditionally, household travel surveys have been collected to be used in modelling travel behaviour. In other words, travel demand modelling, and analysing and/or managing the operation of the transport network, require the availability of detailed information of several types with agents playing a role in generating trips using the transport network. These agents include but are not limited to: individuals, households, vehicles and firms for each of which information about their attributes should be collected, collated, processed and analysed.

Data is generally a valuable product which exhausts a large proportion of the financial resources provided for planning and operating the transport system. As a result, not necessarily all metropolitan areas can afford to collect data on a monthly or yearly basis. This has resulted in the emergence of innovative approaches to temporally and/or spatially transferring data and models [1] or indirectly importing the required data from other readily accessible data sources [1, 2].

Data for demand modelling has been collected using two major methods. These are: i) revealed preference (RP) surveys and ii) stated preference (SP) surveys. Each of these two major methods is used to collect data about a) the household/individual travel diary [3], b) the attitudes or opinions of people about the system and service [4], and c) the number of agents (people or vehicles) using the transport system [5]. Conventional data collection techniques for a and b include face-to-face, telephone, mail-out-mail-back, web-based and on-board (in transit, for example) surveying methods. Count data (c) has been traditionally collected using roadside, GPS, on-board and smart card techniques. The significantly large cost associated with the data collection methods for data types of a and b does not require further discussion as the average cost of one complete household travel survey is more than $200 [6]. As a result, technology has been employed to collect household travel survey data (or even count data) in a cost effective manner. For example, the capacity of web-based surveys (trip planning apps), social networking sites or applications, smart phones (accelerometers) and personal health sensors has been explored. Nonetheless, the practical inherent capacity of these emerging technology-based methods is yet to be explored.

A significant source of data which has been barely considered, along with its capacity for providing household travel information, is the data provided by social media platforms such as Twitter. The main challenge in using such data sources is the significant noise that exists in them which means advanced text mining, linguistic techniques and data mining techniques are required to extract the information that can be related to human travel, which has been discussed in [7, 8]

The literature of travel survey methods has already started the discussion about how smart phones and trip planning apps [9] can be developed to collect travel information with minimum distraction to travellers [10]. Nonetheless, the literature is quite scant when it comes to the application of social media data courses in extracting travel and socio-demographic information.

Generally, the cost of obtaining such social media data is trivial. But processing such massive databases to extract travel information is a challenging task, especially for attributes such as trip purpose. As a result, the accuracy of the outcome is not expected to be high unless advanced data mining and linguistic techniques are used. Nonetheless, the true potential of these techniques in extracting information from social media data is yet to be explored.

Twitter data (tweets) typically contain normal text, hash-tag(s), and/or check-in data. It is relatively easier to work with check-in and hash-tag data as they are already associated with an event or location. For example, check-in data which includes a location for each tweet is associated with activities happening at that location (e.g. tweets linked to a stadium are more likely to be recreational activities). Thus, when check-in data is used for analysing the destination/origin of the activity, then determining trip purpose is relatively straightforward [2-4]. Similarly, hash-tag (#) messages tend to be associated with an activity, event, location or group. For example, if information about Vivid (Sydney's annual Festival of Light, Music and Ideas) is available, then the destination/origin of the activity is fixed. Further, activity time and duration can be easily extracted.

If check-in data or hash-tag data is not of interest and more general information is used, extracting meaningful information can be challenging. The challenges associated with each of the named travel and land use attributes are discussed below. This discussion is based on the assumption that the Twitter data is geocoded, which is often missed for a significant portion of tweets as it is users (who post tweets) who decide to activate the sharing of their geo-location information and most of them are not willing to do this due to privacy issues.

This study attempts to investigate how social media can be used to facilitate and enhance transportation planning, management and operation. This paper is structured as follows. First, the literature is reviewed with a focus on the application of social network data in the field of transportation. Then a comprehensive framework is presented for using social media data in different domains of travel modelling. Next, data preparation and processing approaches are discussed, followed by a detailed discussion about tourist behaviour and trip purpose identification techniques. Finally, a summary is provided and future directions discussed.

2. LITERATURE REVIEW: Harnessing the Potential of Social Media for Travel Demand Analysis

Social networking services or sites (also known as Web 2.0 applications) such as Twitter, Foursquare, Facebook, Flicker and YouTube have revolutionised the way information is produced, shared and stored: anyone can provide information, access and comment on the information. These sites facilitate timely information propagation and to larger audiences. Therefore, such web-based services or applications are also referred to as social media. The rapid development of telecommunications technology and devices such as smart phones and the compatibility of such applications have helped to increase the number of users of such services. These are becoming increasingly popular and millions of documents (e.g. text messages, photos and videos) are published and propagated widely every day. For instance, Facebook’s statistics recorded[1] “890 million daily active users on average for December 2014” with 745 million users using mobile

[1] http://newsroom.fb.com/company-info/

devices. Similarly Twitter claimed[2] "500 million Tweets are sent per day" and there are "288 million monthly active users" with 80% of active users using mobile phones. This creates a great opportunity for both private businesses and the public sector to benefit from the amount of free available information provided online and improve their services.

Crowdsourcing social media for disaster or emergency management is one of the widely used examples of using social media data (e.g. Twitter posts) to facilitate response and relief operations by emergency response organisations. The main aim of such studies is to enhance emergency situation awareness using social media [11]. Among the different approaches, some develop tools to track the information provided by social media in an attempt to predict a likely event [12].

2.1 Using Social Media for Travel Detection and Analysis

In the literature there are a few studies that have focused on applications of social media and social network services in transportation. The relevant existing studies can be divided into two categories: (i) evacuation and traffic incidents; and (ii) user activity pattern-based analysis using geo-tagged data.

Among very few existing relevant studies, Gao et al [13] analysed users' social behaviours from a spatial-temporal aspect using location-based data tracked by 'check-in' applications which enables social media users to share the locations of their trips. Later, Gao et al [14] used similar approach to propose a location-based recommendation system based on the temporal properties of user movement which tracked movements using the same "check-in" data. Such approaches facilitate a variety of services such as traffic forecasting, advertising and disaster relief [13].

Finally, using social media to more effectively manage traffic incidents, Fu et al [15] attempted to study the feasibility of detecting traffic incidents from tweets and also proposed a way to manage incidents more effectively based on the extra information that can be obtained from related Twitter data. They only focused on tweets that contained incident related keywords and evaluated their results by comparing them with real-world incident data. They showed that tweets are useful for early incident detection and can be used as an additional source of information for incident management. A similar approach was taken by Mai and Hranac [16] by comparing recorded incidents of the California Highway Patrol with related tweets via visualising the density of incidents and tweets that coincide near the same location. Steur [17] applied a similar approach to highways in Netherland.

Hasan et al [18] conducted research on location-based data collected from social networking sites to study human mobility and activity patterns. They used users' "check-in" data which contains user activity and geo-location information. Hasan et al also [19] used similar data to extract the weekly activity patterns of individuals and user-specific activity patterns.

2.2 Using Location Based Social Media Data to Study Tourists

The literature is relatively recent when it comes to studies investigating tourist behaviour using social media data. These studies are heavily dominated by research that looks at Flicker data which is geo-tagged and thus the text mining effort is minimal. Xiang and Gretzel [20] studied the role of social media in search engine results when a travel related topic is searched. Girardin et al [21] explored the spatial and temporal variation of the behaviour of tourists by analysing geo-tagged photos shared on Flicker. Popescu and Grefenstette [22] introduced a personalised recommendation system based on users' historical data available on social media and particularly geo-tagged Flicker data. Majid et al [23] introduced a platform to predict tourists visiting preferable places in a new place by analyzing their travel experiences posted on social media sites by sharing geo-tagged photos on Flicker. Sun et al [24] used geo-tagged Flicker photos to study the spatiotemporal behaviour of tourists' accommodation in Vienna, Austria by using the kernel density estimation method. They analysed nearly 245,000 geo-tagged photos with the "Vienna" tag and then filtered them by performing a keyword tag search looking for words related to tourist accommodations. Pozdnoukhov and Kaiser [25] applied LDA to geotagged Twitter data in Ireland to illustrate the spatial-temporal pattern of topics founded in the tweets. Similarly Kling and Pozdnoukhov [26] discovered location based social media obtained from Twitter and Foursquare to explore temporal and spatial topics.

De Choudhury et al [27] developed a method for automatically constructing travel itineraries using geo-tagged Flicker photos. They evaluated itineraries generated for Barcelona, London, New York, Paris and San Francisco by comparing them with bus sightseeing itineraries and crowd source opinions. Ichimura and Kamada [28] developed an Android application to collect tourist subjective data from visited sightseeing spots and automatically analysed it with the Integrated Growing Hierarchical self-organising map.

3. FRAMEWORK

This section presents a detailed discussion on how Twitter data can be used to complement travel demand data sources at a disaggregate level. At the aggregate level, Lee et al [29, 30] have started interesting research directions by looking at the capacity of Twitter data in generating origin destination tables.

Trip Purpose

Trip purpose is one of the most essential attributes in travel demand modelling. Extracting information about the purpose of the activity from the text of tweets is a challenging task. It requires linguistic mining techniques which are still in their infancy although some techniques such as the Latent Dirichlet Allocation (DLA) [31] method is used widely in the literature [32]. The complication is related to the fact that one person may tweet about a restaurant without necessarily meaning to have an outdoor recreational eating activity. When the tweet is combined with the geolocation of the correspondent, it can facilitate extracting the purpose of the trip. The first step in understanding how tweets can be interpreted to determine the purpose of a potential trip is to build a data dictionary. Based on words used in tweets and their co-occurrence with different events or locations, an activity purpose can be determined. As well as looking at the words used in a tweet, combinations of words in sentences should also be looked at (conceptually mining the text). This leads to further advanced linguistic mining techniques, which are not yet used in search engines such as Google.

[2] https://about.twitter.com/company

Departure Time and Activity Location

Given the fact that tweets have a time tag, the process of determining departure time is made easier. The only complication relates to identifying tweets that are related to an activity. In other words, all tweets are related to an activity but the activity might be an indoor activity which is of secondary importance. Using Twitter data to develop understanding about in-home activities is a topic that will be further discussed later.

If an activity is identified as being associated with a tweet, based on the text of the tweet, it can be determined whether the activity has happened, is happening or is scheduled to happen in future. Determining whether an activity precedes or succeeds a tweet can assist in determining the time of an activity and the departure time.

Similarly, the activity location can be determined by the associated geo-location (geo-tagged) coordinates, if provided, and sometimes by looking at the text of the tweet. Unlike check-in (or in some cases, hash-tag) data for which the location of the activity can be easily determined, the geo-location coming with tweets does not necessarily imply the location of the activity, because the respondent might have tweeted before the activity happened or even after the fact. As a result, a combination of GIS methods (to link land use data to the location of tweets) and data mining is required to determine the location of an activity that is related to a tweet.

Travel Route, Activity Duration and Traffic Condition

If nothing is noted in the tweet text about the route or duration of the activity, extracting the information for these travel attributes requires multiple tweets to be jointly considered in order to determine a chain of tweets. This is possible specifically for people who frequently tweet as the record of their tweets operates like a GPS tracker which is also associated with notes about each location or the preceding and succeeding point.

Similarly, traffic conditions and travel time on different links can be determined by looking at travellers tweeting while travelling, in the same way that GPS information is used. The significant advantage of Twitter data to GPS data is that each point is accompanied by a note which may contain more information about the traffic condition.

Mode of Transport and Party Composition

Similar to trip purpose, the mode of transport and party composition could be determined using linguistic mining approaches. Nonetheless, constructing a dictionary for this purpose is not as complicated.

Socio-demographic Attributes

Using data mining techniques, it is possible to determine the location of the home, workplace and school of users in the same way that activity locations are determined as it is more likely that many tweets are posted from home, school and work. Keywords related to home, work and school are expected to be present more in tweets posted from these locations. Therefore by looking at a history of tweets by a user, it becomes possible to figure out these three important locations.

A panel data for job relocation history is costly, requiring a long term project to follow people and observe their relocation pattern. Alternatively, retrospective data can be collected for this purpose. The freely available data from social networking services such as LinkedIn, in which people retrospectively report their job/school relocation pattern, can be used for this purpose. The significance of using LinkedIn data pertains to the accuracy of information about job/school location, relocation timing and type. Therefore, by mining the Twitter and LinkedIn data, some demographic information such as gender, estimated income, social network type and age can be determined.

4. DATA

In this section, we showcase our approach by identifying and analysing the trips of people who visited the Sydney (Australia) metropolitan area during a period of about a month and also short- or long-distance trips (travels) of Sydney residents.

The ideal approach would be to investigate online data that have geo-location (geo-tagged) data attached. Twitter and some other social media platforms have this capability by including the latitude and longitude of the location at which a user posts information (which ideally is the same as the location for the activity the user is sharing information about). But this function is not active by default and users have the liberty to activate it or not. Privacy issues are the main concern for users to opt out this service.

Data for this study has been extracted from Twitter, using 'Twitter Search API', from 11 May 2015 to 5 June 2015 (inclusive) by requesting the tweets within Sydney metropolitan area, setting the city of Parramatta and the radius of 25 miles as the geo-location in the query. Since Twitter API returns only the most recent data (of the past 8 to 9 days), the data collection has been conducted four times. The data has been merged and redundant data has been removed. In total 85,740 tweets with geographical coordinates (geo-location or geo-tagged) data and slightly less than 11,000 users were extracted and stored in our dataset for further analysis. Henceforth, we will refer to this dataset as 'Sydney_GeoLoc_DB'.

Figure 3 depicts the frequency of the geo-tagged tweets and the number of unique geo-locations (where the tweets have been posted) during the data collection period. The number of unique geo-locations is much smaller as usually the tweets are posted from one or more repeated areas (e.g. home, office or school). The figure also illustrates that more tweets and locations have been posted during weekends, as expected. The counter-intuitive relatively high number of geo-tagged tweets posted from 7 to 8 May reveals a potential special event. The Sydney event calendar indicated that this was the last weekend of the Sydney Vivid festival, which also supported by the high number of tweets using the #sydneyvivid or alternative hash-tags.

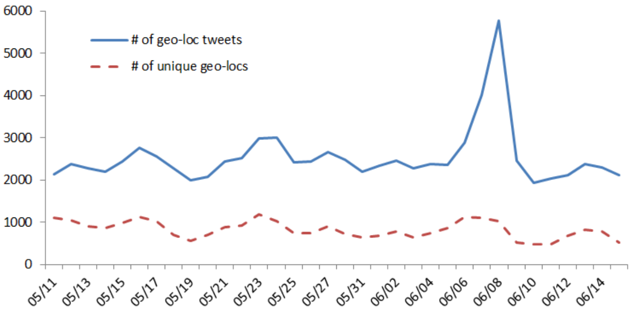


Figure 3. The frequency of geo-tagged tweets and unique geo-locations over time

Figure 4 shows the number of geo-tagged tweets on different areas of Sydney metropolitan area. The figure shows high numbers at several city centres, with the main areas being Sydney CBD (Central Business District), Parramatta City Centre and the eastern beaches.

5. METHODOLOGY AND RESULT

5.1 Identifying Sydney Residents Visiting Other Cities

In order to identify Sydney residents and also visitors we needed to have access to geo-tagged tweets. However, there is not an easy way to identify such users (to our knowledge). Thus, in order to find users who are providing geo-location, we have used Twitter Search API to search tweets in the Sydney metropolitan area.

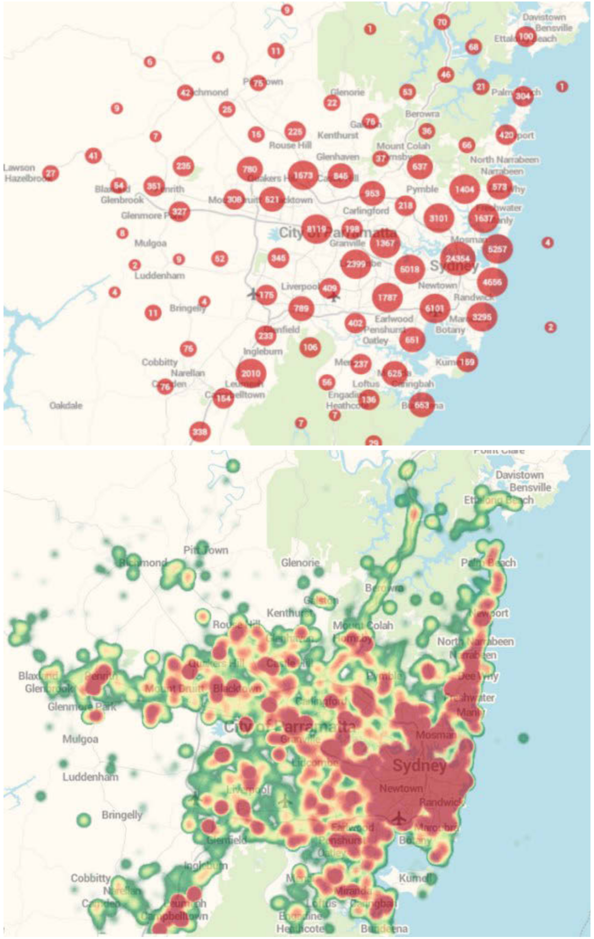


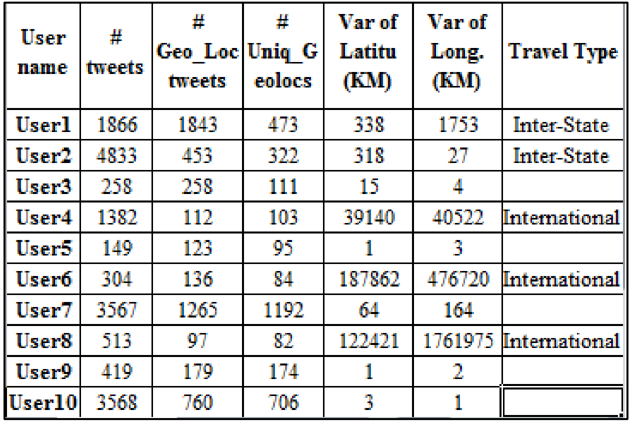
Figure 4. Mapping 85,000 geo-tagged tweets in the Sydney metropolitan area

To analyse Sydney residents' trips, we needed to identify the residents first. So, we kept users having at least 10 unique geo-tagged tweets in at least three phases of the data collection phases (during the construction of 'Sydney_GeoLoc_DB' dataset). Applying these criteria on the extracted dataset revealed about 400 users as potential Sydney residents. Then, we checked the users' profile location, the self-reported location Twitter. After removing companies or individual advertisers we ended up with about 310 Sydney residents actively using Twitter who also often share their location. Figure 4 illustrates the number of tweets posted by all the users in our database, in different areas of Sydney metropolitan area. The bottom figure (heat map) shows a more scattered distribution. Both figures clearly confirm that Twitter data well demonstrates the reality, highlighting more activities in main areas (city centres and coastal areas).

In separate attempts, we used different queries to extract all the tweets (irrespective of having geo-location information) of all the 310 identified Sydney residents. This helped to make a second database (Sydney residents' tweets) containing about 155,000 tweets from which only 26,000 were geo-tagged. Henceforth, we will refer to this dataset as 'Sydney_Resident_DB'.

Such a dataset enabled us to calculate and compare the number of tweets, the number of geo-tagged tweets, and also the number of unique geo-locations – as it is possible to geo-tag many tweets from a similar location such as home or office, as shown in Table 1. More unique geo-tagged tweets by a user reflect trips between those places. In addition, we calculated and used the variance of latitudes and longitude of the geo-coordinates (provided for each geo-tagged tweet) for each user to understand the distance of their trips. Thus, the variance for less than 100 km clearly shows short trips in the city while above 100 km and less than 1000–2000 km demonstrates interstate travels and more than 2000 km reveals international travels. Table 1 illustrates the figures for a sample of 10 users from the 'Sydney_Resident_DB'.

Table 1. Trip distance extraction and comparison for sample of Sydney residents



User name	# tweets	# Geo_Loc tweets	# Uniq_Geolocs	Var of Latitu (KM)	Var of Long. (KM)	Travel Type
User1	1866	1843	473	338	1753	Inter-State
User2	4833	453	322	318	27	Inter-State
User3	258	258	111	15	4	
User4	1382	112	103	39140	40522	International
User5	149	123	95	1	3	
User6	304	136	84	187862	476720	International
User7	3567	1265	1192	64	164	
User8	513	97	82	122421	1761975	International
User9	419	179	174	1	2	
User10	3568	760	706	3	1	

5.2 Identifying Sydney Tourists

Using the first dataset, 'Sydney_GeoLoc_DB', we were able to extract the users in each data collection phase and then compare the lists with each other as well as with the other dataset 'Sydney_Resident_DB' to identify the (potential) tourists (travellers).

In each phase of data collection between 8,500 and 11,000 users (with at least one geo-tagged tweet) were reported. The users who had been in only one (or two) phase of the data collection (out of four phases) were considered as tourists. We also applied a threshold of minimum 9 unique geo-tagged tweets at each phase to focus on more active users. Applying these filters, we identified about 580 users, who had posted about 14,700 tweets, as travellers to Sydney. Figure 5 illustrates the locations they visited during their stay in Sydney.

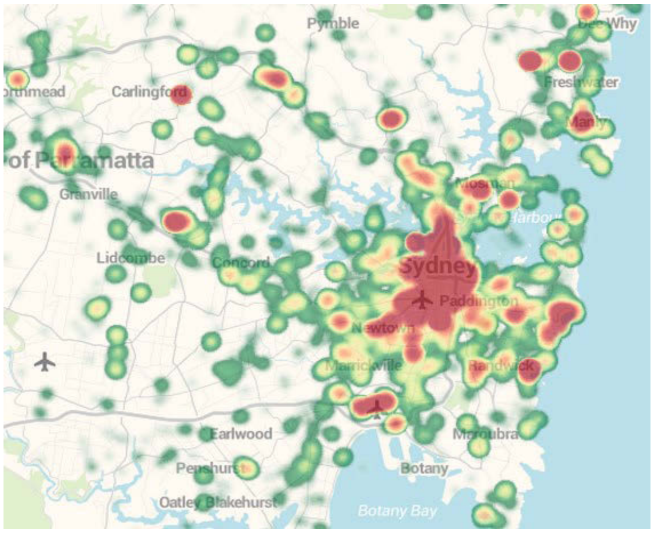


Figure 5. Sydney tourist visits (based on geo-tagged tweets) in Sydney metropolitan area

Table 2. Land Use of Sydney Tourists versus Residents

Land Use Category	Tourists	Residents
Conservation Area	**107**	**87**
Cultural heritage site	21	4
National park	85	73
River & Drainage System	**1,088**	**503**
River, creek or …	875	386
Marina	192	60
Special Categories	**108**	**104**
Beach	108	104
Transport & Other Corridors	**4,206**	**5,785**
Aerodrome/airport	269	154
Railway	323	540
Road or road reserve	3,613	5,065
Urban	**9,068**	**15,034**
Industrial/commercial	3,248	3,588
Residential	2,786	8,571
Urban recreation	1,721	1,516
Rural residential	59	57
Surf club and/or coastal facil.	22	13
Tourist development	38	16
University or other tertiary inst.	279	364
Government and private facil.	886	883
Golf Course	17	6
Others	**129**	**383**
Total	**14,706**	**21,896**

In a further analysis, we compared the attributes of the locations tourists visited in Sydney with the locations visited by the identified Sydney residents. To do this, using a GIS exercise, we mapped the geo-coordinates of the tweets with the land use data provided by NSW Land and Property Information. Table 2 shows the number of tweets for each category of land use for both the 580 identified tourists and 310 residents. The numbers in the table represent each community. As expected, tourists are tweeting more on 'transport corridors', 'conservation areas', 'River and Drainage System' or 'beaches' while the residents tweet more from their 'residential' areas.

It is important to note that each parcel in the spatial data is associated with one dominant land use type. As a result, for mixed land use types, especially in the CBD area, no distinction between shopping malls, government facilities and residential units is provided.

5.3 Activity Purpose Identification

5.3.1 Methodology (Modified LDA)

Latent Dirichlet allocation (LDA) [31] is a hierarchical Bayesian based approach for finding similarities among categorical variables. As discussed earlier, LDA has been used in the literature for analysing social media content (e.g. tweets). LDA is a latent space model which identifies correlations between words in text corpora to explore topics and potential ways of classifying texts. In this section we introduce a modified version of LDA for the purpose of extracting a trip. Unlike in similar studies, we mainly focused on the content of tweets rather than check-in data [13], geo-tagged data [19, 33] or sentiment analyses [15, 34]. First, we built our own dictionary, including a set of unique words, 'Sydney_Resident_DB'. At this stage prepositions and symbols were not considered. Second, based on the word counts, the around 400 words that had been repeated in the database at least 20 times were shortlisted. For example, words like "I'm", "Sydney" and "restaurant" were respectively used 3009, 1215 and 233 times in the database that we selected for further analysis. Nearly 17,000 words were not included in the analysis because they were slang expressions, prepositions, symbols or had been repeated fewer than 20 times in the database. Third, LDA was utilised to cluster the 400 selected words. We came up with almost 100 word clusters. Fourth, for each cluster we identified the top 3 words, that is those with the highest number of counts which are also in the pool of 400 words listed. Fifth, the correlation between these top 3 words and other words, that are also repeated frequently but not as many times as top 3 words, were defined based on the number of co-occurrences. Sixth, by looking at the top 3 words in each cluster and considering the other correlations among words an activity tag was able to be assigned to that cluster. Activity tags can be one of the following:

- Shopping
- Entertainment
- Eating
- Work
- Social
- Study

Finally, each tweet was checked against all the clusters and if there was an appropriate level of similarity then the cluster's tag was assigned to the tweets. The tagging process was not limited to one tag. So multiple tags could be associated with each tweet.

At the moment, this process is computationally extensive, However, this part of the research is ongoing and needs further improvement when it comes to the efficiency of the LDA method in determining travel attributes from Twitter data.

5.3.2 Results

The initial results, as summarised in Table 3, show the trip purposes as identified for part of 'Sydney_Resident_DB' for which meaningful text was in hand. Although this indicates we are able to tag fewer than 20% of all tweets, it also indicates the potential for using social media for transportation purposes at the highly disaggregate level which is one of the major contributions of this work. The other important finding that can be seen in Table 3 is that social media, and particularly Twitter, is used for entertainment, eating and social activities rather than simply work and study activities for people who post tweets at multiple locations (transport engineers call them travellers).

Table 3. Trip purpose of the tweets

Trip Purpose	Tagged tweets
Shopping	2%
Entertainment	47%
Eating	27%
Work	3%
Social	18%
Study	3%

Figure 6 presents the spatial distribution of tagged activities around the Sydney metropolitan area which has a dense accumulation of entertainment activities around the CBD and social activities in certain suburbs. Further, it can be observed from the map below that study activities are happening in and around universities such as UNSW and the University of Sydney.

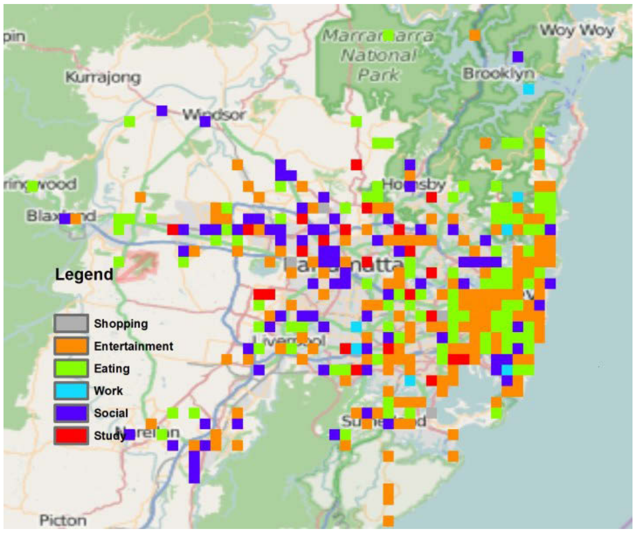


Figure 6. Distribution of tweets based on extracted trip purpose

6. SUMMARY AND CONCLUSION

This paper presented a proof of concept about the potential for Twitter data to be used in analysing the individual level travel behaviour of users. A framework was initially discussed in which the applicability of Twitter data in determining travel characteristics for various travel attributes at the highly disaggregate level was illustrated. This framework paves the way toward more applications of Twitter and other social media data for transport planning, management and operation purposes.

Following the discussion about the framework, three components of the framework are analysed in this paper with a special focus on tourism. Longitudinal data was obtained from Twitter for this study which enabled us to identify tourists (visitors) and study their movements around Sydney. Several data fusion techniques were employed to fill in the missing information, such as the city/country of origin of tourists. Further, the location of tweets were mapped to fine resolution using the land use data of the City of Sydney in order to analyse the destination choice behaviour of tourists versus regular Sydney residents. Several intuitive findings were obtained by studying the movement of tourists around the city, such as the high density of Twitter posts at facilities which represent the recreational activities happening at those location.

To further elaborate the feasibility of using Twitter data for travel behaviour analysis at the disaggregate level of individuals, this paper presented the preliminary results of an endeavour for determining the activity purpose associated with each tweet using an advanced text mining technique, LDA. It was found that tweets are mainly associated with recreational activities and the spatial distribution of these activities matches the distribution of facilities around the city. This finding further admits the usefulness of Twitter data for analysing the behaviour of tourists in tourist attracting cities like Sydney.

Other than the above-mentioned advantages of using social media data, three other aspects of transport planning and travel demand modelling can be discussed that are more related to the activity-based modelling scheme. We plan to consider these in our future works:

1. In-home activity data: The area of travel demand modelling to obtain data about in-home activities of people presents challenges. This is important to travel demand modellers, specifically activity-based modellers, because there is a trade-off between the hours people spend for some types of activity like eating in the home and out of the home. If the activity is scheduled to happen at home, one out-of-home activity is cancelled, which results in fewer trips on the transport network, which is of great importance to travel demand modellers and planners.

2. Tour formation: tour-based models are among advanced demand modelling approaches which require collecting information about trips forming a tour of activities typically starting from home and ending at home. Twitter users often provide information about their daily activities which can be mined to extract information about the location, time and purpose of different activities, especially if it is linked with land use data. Using Twitter data for modelling tour formation behaviour can significantly complement the models that are developed using household travel surveys.

3. Future activities: When the Twitter data is mined using linguistic techniques, it becomes possible to forecast potential future activities. In other words, if future tense is used in a tweet, and a location is stated about an activity to happen soon in the future (in less than a week), this can imply that the person is likely to be at that location at a time to be determined. When a model processes the contents of tweets and approximates the number of trips likely to happen in a short run in future, the operation and management of the transport system can be facilitated. This has a significant impact on evacuation management and the management of any disruptions to the network as a result of an accident or other large event.

7. REFERENCES

[1] Rashidi, T., J. Auld, and A. Mohammadian, Effectiveness of Bayesian Updating Attributes in Data Transferability Applications. Transportation Research Record: Journal of the Transportation Research Board, 2013(2344): p. 1-9.

[2] Miller, E., et al., A Framework for Urban Passenger Data Collection, , in 10th International Conference on Transport Survey Methods. 2014: Leura, Australia.

[3] Rashidi, T., A. Mohammadian, and Y. Zhang, Effect of Variation in Household Sociodemographics, Lifestyles, and Built Environment on Travel Behavior. Transportation Research Record: Journal of the Transportation Research Board, 2010(2156): p. 64-72.

[4] Beirão, G. and J.S. Cabral, Understanding attitudes towards public transport and private car: A qualitative study. Transport policy, 2007. 14(6): p. 478-489.

[5] Francis, R.C., et al., Object tracking and management system and method using radio-frequency identification tags. 2003, Google Patents.

[6] Zhang, Y. and A. Mohammadian, Bayesian updating of transferred household travel data. Transportation Research Record: Journal of the Transportation Research Board, 2008(2049): p. 111-118.

[7] Cramer, H., M. Rost, and L.E. Holmquist. Performing a check-in: emerging practices, norms and'conflicts' in location-sharing using foursquare. in Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services. 2011. ACM.

[8] Maghrebi, M., et al., Complementing Travel Diary Surveys with Twitter Data: Application of Text Mining Techniques on Activity Location, Type and Time, in 17th International Conference on Intelligent Transportation Systems (ITSC). 2015, IEEE: Las Palmas, Spain.

[9] Williams, S. and E. Currid-Halkett, Industry in Motion: Using Smart Phones to Explore the Spatial Network of the Garment Industry in New York City. PloS one, 2014. 9(2): p. e86165.

[10] Byon, Y.-J., B. Abdulhai, and A. Shalaby, Real-time transportation mode detection via tracking global positioning system mobile devices. Journal of Intelligent Transportation Systems, 2009. 13(4): p. 161-170.

[11] Yin, J., et al., Using social media to enhance emergency situation awareness. IEEE Intelligent Systems, 2012. 27(6): p. 52-59.

[12] Cameron, M.A., et al. Emergency situation awareness from twitter for crisis management. in Proceedings of the 21st international conference companion on World Wide Web. 2012.

[13] Gao, H., J. Tang, and H. Liu. Exploring Social-Historical Ties on Location-Based Social Networks. in ICWSM. 2012.

[14] Gao, H., et al. Exploring temporal effects for location recommendation on location-based social networks. in Proceedings of the 7th ACM conference on Recommender systems. 2013.

[15] Fu, K., R. Nune, and J.X. Tao. Social Media Data Analysis for Traffic Incident Detection and Management. in Transportation Research Board 94th Annual Meeting. 2015.

[16] Mai, E. and R. Hranac. Twitter Interactions as a Data Source for Transportation Incidents. in Proc. Transportation Research Board 92nd Ann. Meeting. 2013.

[17] Steur, R., Twitter as a spatio-temporal source for incident management. 2015.

[18] Hasan, S., X. Zhan, and S.V. Ukkusuri. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. in Proceedings of the 2nd ACM SIGKDD international workshop on urban computing. 2013. ACM.

[19] Hasan, S. and S.V. Ukkusuri, Urban activity pattern classification using topic models from online geo-location data. Transportation Research Part C: Emerging Technologies, 2014. 44: p. 363-381.

[20] Xiang, Z. and U. Gretzel, Role of social media in online travel information search. Tourism management, 2010. 31(2): p. 179-188.

[21] Girardin, F., et al., Digital footprinting: Uncovering tourists with user-generated content. Pervasive Computing, IEEE, 2008. 7(4): p. 36-43.

[22] Popescu, A. and G. Grefenstette. Mining social media to create personalized recommendations for tourist visits. in Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications. 2011.

[23] Majid, A., et al., A context-aware personalized travel recommendation system based on geotagged social media data mining. International Journal of Geographical Information Science, 2013. 27(4): p. 662-684.

[24] Sun, Y., et al., Analyzing human activities through volunteered geographic information: Using Flickr to analyze spatial and temporal pattern of tourist accommodation, in Progress in Location-Based Services. 2013, Springer. p. 57-69.

[25] Pozdnoukhov, A. and C. Kaiser. Space-time dynamics of topics in streaming text. in Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based social networks. 2011.

[26] Kling, F. and A. Pozdnoukhov. When a city tells a story: urban topic analysis. in Proceedings of the 20th International Conference on Advances in Geographic Information Systems. 2012.

[27] De Choudhury, M., et al. Automatic construction of travel itineraries using social breadcrumbs. in Proceedings of the 21st ACM conference on Hypertext and hypermedia. 2010.

[28] Ichimura, T. and S. Kamada. A generation method of filtering rules of Twitter via smartphone based Participatory Sensing system for tourist by interactive GHSOM and C4. 5. in Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on. 2012.

[29] Lee, J.H., A.W. Davis, and K.G. Goulias, Activity Space Estimation with Longitudinal Observations of Social Media Data, in Paper submitted for presentation at the 95th Annual Meeting of the Transportation Research Board. Washington, D.C., January 10-14, 2016.

[30] Lee, J.H., et al., Can Twitter data be used to validate travel demand models?, in 14th International Conference on Travel Behaviour Research. 2015: Windsor, UK.

[31] Blei, D.M., A.Y. Ng, and M.I. Jordan, Latent dirichlet allocation. the Journal of machine Learning research, 2003. 3: p. 993-1022.

[32] Coffey, C. and A. Pozdnoukhov. Temporal decomposition and semantic enrichment of mobility flows. in Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-Based Social Networks. 2013. ACM.

[33] Hasan, S. and S.V. Ukkusuri, Social contagion process in informal warning networks to understand evacuation timing behavior. Journal of Public Health Management and Practice, 2013. 19: p. S68-S69.

[34] Kaigo, M., Social media usage during disasters and social capital: Twitter and the Great East Japan earthquake. Keio Communication Review, 2012. 34: p. 19-35.