2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)
Windsor Oceanico Hotel, Rio de Janeiro, Brazil, November 1-4, 2016

Transportation Application of Social Media: Travel Mode Extraction

Mojtaba Maghrebi [1,2*], Alireza Abbasi [3], S. Travis Waller [1]

Abstract- **At the present time, social media is not only used for connecting people in a virtual environment, but is also considered as a reach source of information for organizations and public service agencies to facilitate their policy and decision making processes. As an example, such crowdsourced data can be considered as a complementary source for analyzing people choices. In this study, we attempt to show how social media data can be used (and utilized) in order to extract travel mode choice which can be used as complementary source of information to improve traditional costly methods such as House Travel Surveys (HTS). The contents of Twitter data posted in Melbourne metropolitan areas have been analyzed to determine travel mode choices information. The results show, walking and driving modes are the most frequent travel modes extracted from Twitter data while public mode of transportations such as bus and taxi are rarely detected. Future research is required to extend this approach by considering and validating socio-demographic metrics of social media users so as to utilize social media data as complementing source of information for HTS.**

I. Introduction and Background

Around eighty percent of American use social media and near two third of the global internet population visits social networks [1]. Social media provides a chance to have and share on-time or recently updated data which makes it more attractive. Also, the cost for collecting this often freely available data is much cheaper than traditional data collection process. So, this has extended the use of social media from only connecting people to a rich and massive data source to study human interactions and behaviour in different industries such as health care [2-4], marketing [5-7], food industry [8, 9], emergency management [10-13] and transportation [14-16] to facilitate their operation or management.

A social media post and particularly a tweet (a post in Twitter) typically contains time, date and text and may has geographical coordinates (geo-tagged). In the literature, mainly time, date and geographical coordinates are used for spatial-temporal analyses and content of tweets has not received attentions as other components of a post due to the complexity of text mining and text analyzing techniques.

The main challenge which makes utilizing social media data complicated is the huge size of available data which often needs more advanced computing techniques to collect and store such a massive data and shorten the processing time.

II. Literature Review

For instance, Fu et al [17] investigated the feasibility of incident detection using information from the people posts in Twitter in order to facilitate the effective management of the emergency situations. Their research methodology is limited to searching keywords that possibly are related to the incidents social media's users involved or observed. Similarly, Mai and Hranac [18] proposed social media data can be used to alert an incident. They compared the Twitter posts (tweets) near the roads with the real incident reports by California Highway Patrol and showed that when the density of tweets increased most likely an accident has happened. Similar approach and results obtained by Steur [19] but from Twitter data in Netherland.

In the past couple of years a growing body of literature has been devoted to transportation applications of social media [20]. For example, Gao et al [21] studied the people behavior by applying spatial-temporal analyses on check-in data. They then proposed a location-based recommendation system by considering consequence of check-ins posted in

[1] Research Centre for Integrated Transport Innovation, School of Civil and Environmental Engineering, UNSW Australia, Sydney NSW 2052, Australia

[2] Department of Civil Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.

[3] School of Engineering and Information Technology, The University of New South Wales (UNSW Australia), Canberra, ACT, 2610, Australia
{maghrebi, a.abbasi, s.waller)@unsw.edu.au

978-1-5090-1889-5/16/$31.00 ©2016 IEEE

Foursquare [22]. It is much easier to deal with check-in and hash-tag data as they are already associated with a special event or location. Check-in data can be used for examination of the destination/origin of the activity as well as easily detecting purpose of trip. Likewise, hash-tag (#) posts are linked with an event, activity, location or a social campaign. Exploring the text content of tweets are challenging than only using numerical values of data, time and coordination. Among very few, Abbasi et al. [14] used geo-tagged Twitter data of the individuals in Sydney, Australia in order to detect their movements and activity purposes by using the land use data and analyzing the content of the tweets proposing and extended latent Dirichlet allocation (LDA) technique. They also proposed a method to distinguish travelers (tourists) and residents and then compared the movement and activity purpose of the two groups.

To analyze human mobility patterns, social media platform which provide check-in data such as Foursquare or Yelp has been used widely. For instance Hassan et al. [23] used Foursquare check-in data to study people weekly activity patterns. Girardin et al [24] explored the geo-tagged photos shared on Flicker to study the tourist movement behavior. Majid et al [25] proposed a method to find tourists preferable places by analyzing shared geo-tagged photos on Flicker. Similarly, Sun et al [26] used geo-tagged photos in Flicker and applied kernel density estimation method associated with keyword search to determine tourist preferred accommodations in Vienna, Austria.

Pozdnoukhov and Kaiser [27] looked at the contents of Twitter posts to figure out the spatial-temporal pattern of topics in Ireland. De Choudhury et al [28] proposed a model to automatically build travel itineraries using geo-tagged Flicker photos and validated their model by comparing their results against itineraries of bus sightseeing in Barcelona, London, New York, Paris and San Francisco. Ichimura and Kamada [29] developed an Android application to collect tourist subjective data from visited sightseeing spots and automatically analyzed it with the Integrated Growing Hierarchical self-organizing map.

One of the areas which has received very few attention is how to use social media data to extract travel mode effectively and use it for travel demand analysis. This research attempt to fill this gap by using a sample data of nearly 300,000 tweets collected for about six month for Melbourne metropolitan areas to show to what extent social media data (and in particular Twitter) can help in detecting people travel mode and how this can be used as a complementary source of information for HTS.

III. Methodology

A. Data Collection and Description

Data for this study has been retrieved from Twitter between 1 Nov. 2015 and 19 April 2016 (inclusive) over almost 23 weeks period by requesting the tweets within Melbourne metropolitan area. Since Twitter API returns only the most recent data (of the past 8 to 9 days), data gathering has been conducted on a daily basis. After merging the data and removing the redundant records, a total of 293,628 geo-tagged tweets (i.e., tweets with geographical coordinates' information) have been stored in our dataset for further analysis. Further analysis revealed 25,463 unique user accounts have been posted these geo-tagged tweets which almost 10,000 user accounts have only one geo-tagged tweet. Please note that although geo-tagged data is not required for extracting travel mode data but we have used this data which has been prepared for a separate study [30] to analyze human mobility patterns. To have more insight about the collected database, some features of data are considered deeply.

Figure 1 illustrates the density of posted tweets across the Melbourne metropolitan areas where can see respectively from Melbourne CBD (Central Business District), Geelong and Mornington peninsula more Twitter activities were observed.

Figure 2 shows the frequency of the geo-tagged tweets during the data collection period. The vertical lines separate weeks (starting from Sunday). As can be seen the frequency of geo-tagged tweets in Melbourne metropolitan area is between 1500 and 2000 tweets per day on average with four larger number of posts in weekends of week 2 and 3 of the Dec 2015 and third weeks of Feb and March 2016. The figure shows an overall trend for each week with slight punctuations (a fall at the beginning of the week and sharp increase towards the weekends).

If we aggregated the tweets in weekly horizon, can see that there is no big difference between the weekly slots and with exception of a few weeks the most of the weeks the total numbers of tweets are about the same (Figure 3). Now, the daily frequency of tweets is explored. Figure 4 illustrates the variation of daily frequency cross the entire period of data collection and Figure 5 the total number of tweets in each day of week. By considering the median values can find a steady increase towards the weekends with a slight drop for Sundays. Also generally can summarize that on Saturdays more tweets were posted and on Mondays and Tuesdays least activities observed.

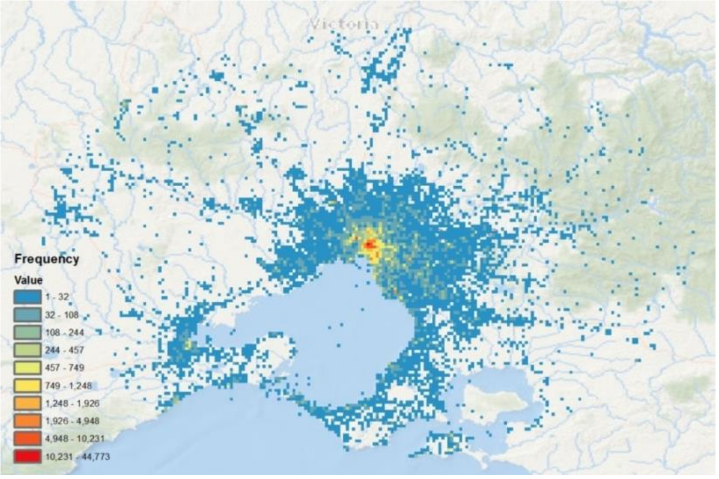

Figure 1. The heat map of geo-tagged tweets in Melbourne per week between Nov 2015 and April 2016

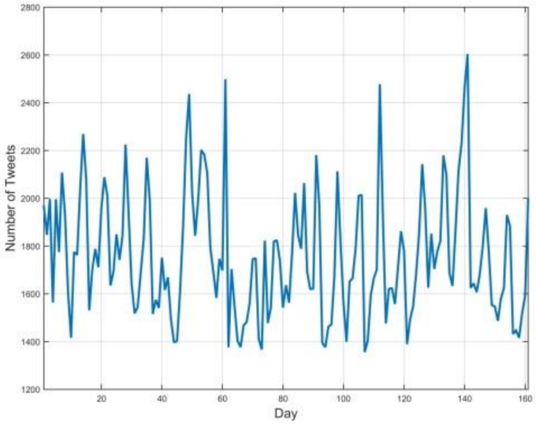

Figure 2. The frequency of the geo-tagged tweets in Melbourne per day between Nov 2015 and April 2016

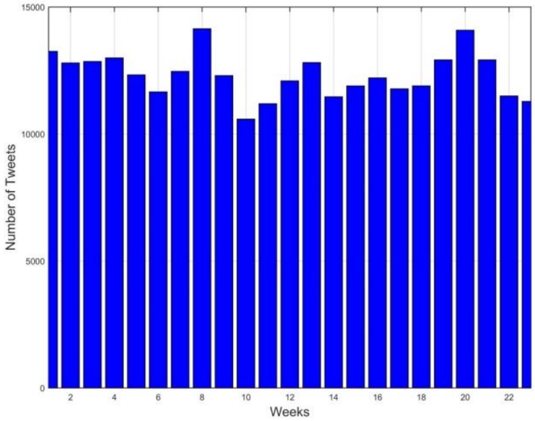

Figure 3. The frequency of the geo-tagged tweets in Melbourne per week between Nov 2015 and April 2016

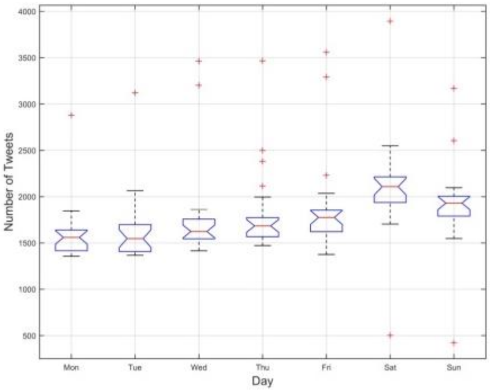

Figure 4. The frequency of the geo-tagged tweets in Melbourne per day between Nov 2015 and April 2016

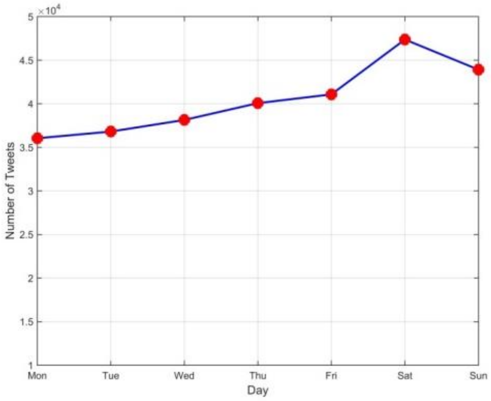

Figure 5. The total number of the geo-tagged tweets in Melbourne per day between Nov 2015 and April 2016 for each day of week

Finally the frequency of tweets in an hourly manner is illustrated in Figure 6. The hourly frequency of geo-tagged tweets for Melbourne, in terms of time of the day, reveals people surprisingly posting tweets (or share their location in the tweets) more frequently during mid-day and mid-night. The most popular time for sharing the locations in Twiter is found to be after mid-night until mid-day which is surprisingly much higher than the afternoon and evening which people are most likely outside their house or work place.

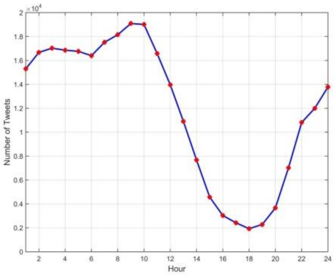

Figure 6. Aggregated frequency of geo-tagged tweets comparing days of a week and hours of a day

B. *Travel Mode Extraction*

In order to identify the travel mode of the tweeter's users in Melbourne metropolitan area we have applied a content analysis technique to find specific words and keywords that most likely were used to express the travel situations associated with the activity described in the contents of tweets.

Twitter does not allow more than 140 characters in a post which is even shorter than normal SMS with 160 characters limits. This issue forces the users to shorten their posts which might cause dropping some part of information that they supposed to share. If the transportation mode is not the main point to be shared then it would be very hard to extract it from the other available source of information included in a tweet.

TABLE 1: TERMS AND VARIATIONS ASSOCIATED WITH TRANSPORTATION MODES

Mode of Transport	Term and variations
Bus	"bus"
Train	"train", "metro"
Taxi	"taxi"
Car	"car", "drive", "driving", "drove", "parking", "traffic"
Bike	"bike", "cycle"
walk	"walk"

Before applying content analyses, the extracted Twitter data was cleaned to avoid from missed values and the tweets with contents rather than English are ignored. Then cleaned geo-tagged tweets are used for content analysis.

Looking for specific keywords in social media data to extract the desirable information has been used in the existing studies [18, 31-33]. Likewise, in this research, we define a set of words and their variations (summarized in Table 1) to extract the following modes of transports recorded in Twitter data: Bus; Train; Taxi; Car (private); Bike; and Walk.

IV. RESULTS AND DISCUSSION

To automate the process a code is developed in MATLAB R2014b and It worth to mention that all the texts are converted to lowercase because of text sensitivity concerns and also to catch all matching tweets. Figure 7 shows the proportion of each mode extracted from the Twitter data. surprisingly, walking mode found the most frequent mode of travel in the data which may related to this fact that the users are more flexible when walking and can easily share their though. After walking mode, respectively driving, train and cycle are most frequent mode of travel while taxi and bus terms are rarely used in the tweets.

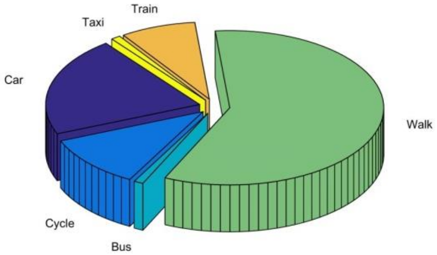

Figure 7: Proportion of travel models extracted from Twitter data

Figure 8 demonstrates the spatial analysis of all tweets associated with travel model in a heat map. The interesting point is that the frequency of tweets related to travel modes is not pretty much similar to frequency of tweets (Figure 1).

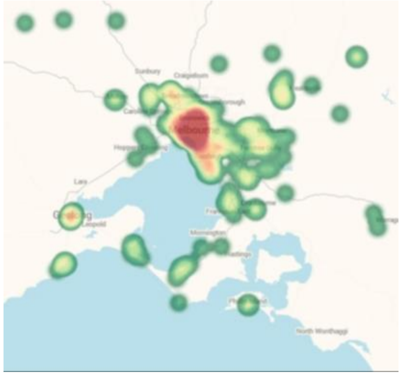

Figure 8: Spatial analyses of all travel modes extracted from twitter data

Figure 9 maps the users' travel mode choice on the Melbourne metropolitan area comparing each mode in sub-figures (a) to (f). The achieved results can be used as complementing source of information if being validated and the socio-demographic parameters of the social media users (e.g., gender, age, income, level of education, race, martial status, etc.) are taken into account. We have not reported the actual numbers and have tried to only present a proof of concept and summarized the aggregated results because these values without validation are questionable. In future works, the obtained results can be compared with zone aggregated house travel survey results to check the level of accuracy and reliability of the available results.

V. Conclusion

Crowdsourced databases such as social media contents can be seriously considered as a rich source of information for solving practical problems. In this paper, we proposed a potential application of social media in transportation context. We extracted mode of travel associated with the tweets. To demonstrate the idea we collected near 300k geo-tagged tweets posted in metropolitan area during 23 weeks between December 2015 and April 2016. After initial data pre-processing and data cleaning, the tweets' contents are analyzed to determine the posts with travel mode choices information. The results show walking and driving modes are the most frequent travel modes extracted from Twitter data while public mode of transportations such as bus and taxi are rarely detected in the available social media database. These results can be used as complementing source of information for House Travel Surveys (HTS) after being validated by considering socio-demographic metrics of social media users.

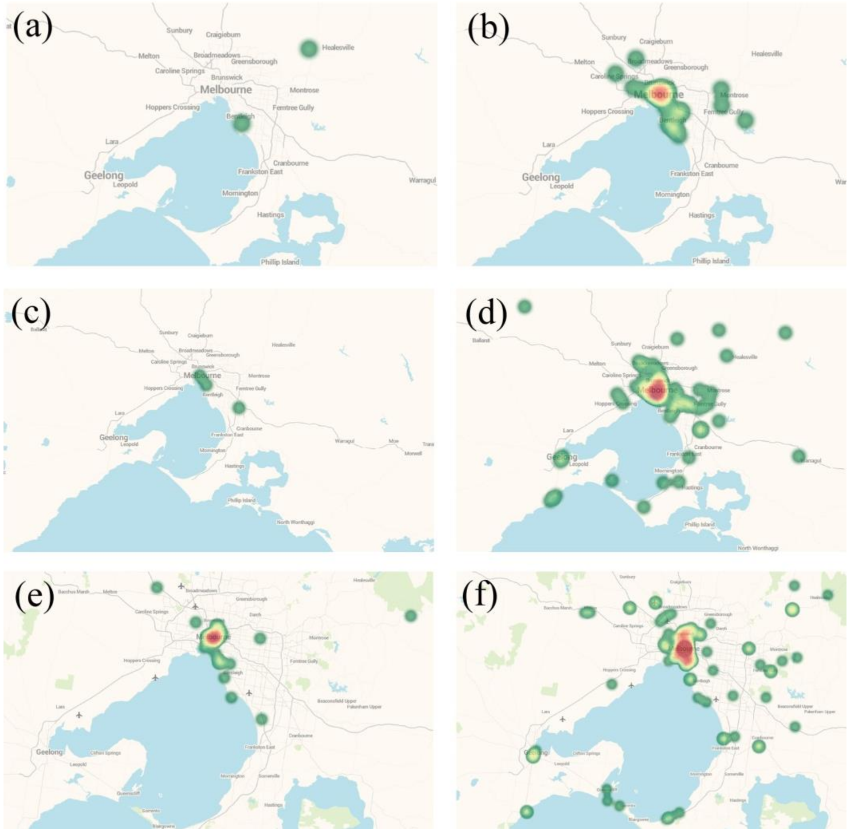

Figure 9: Spatial analyses of each travel modes extracted from twitter data: (a) bus, (b) train, (c) taxi, (d) car, (e) bike and (f) walk.

References

[1] A. Perrin, "Social media usage: 2005-2015 65% of adults now use social networking site–a nearly tenfold jump in the past decade," *Washington, DC: Pew Research Center,* 2015.

[2] R. Thackeray, B. L. Neiger, A. K. Smith, and S. B. Van Wagenen, "Adoption and use of social media among

public health departments," *BMC public health,* vol. 12, p. 1, 2012.

[3] H. Korda and Z. Itani, "Harnessing social media for health promotion and behavior change," *Health promotion practice,* vol. 14, pp. 15-23, 2013.

[4] B. L. Neiger, R. Thackeray, S. A. Van Wagenen, C. L. Hanson, J. H. West, M. D. Barnes*, et al.*, "Use of social media in health promotion purposes, key performance indicators, and evaluation metrics," *Health promotion practice,* vol. 13, pp. 159-164, 2012.

[5] D. Evans, *Social media marketing: the next generation of business engagement*: John Wiley & Sons, 2010.

[6] A. M. Kaplan, "If you love something, let it go mobile: Mobile marketing and mobile social media 4x4," *Business horizons,* vol. 55, pp. 129-139, 2012.

[7] S. A. Khan and R. Bhatti, "Application of social media in marketing of library and information services: A case study from Pakistan," *Webology,* vol. 9, pp. 1-8, 2012.

[8] M. J. Maloni and M. E. Brown, "Corporate social responsibility in the supply chain: an application in the food industry," *Journal of business ethics,* vol. 68, pp. 35-52, 2006.

[9] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," *International Journal of Information Management,* vol. 33, pp. 464-472, 2013.

[10] N. L. Chan and B. D. Guillet, "Investigation of social media marketing: how does the hotel industry in Hong Kong perform in marketing on social media websites?," *Journal of Travel & Tourism Marketing,* vol. 28, pp. 345-368, 2011.

[11] Z. Xiang and U. Gretzel, "Role of social media in online travel information search," *Tourism management,* vol. 31, pp. 179-188, 2010.

[12] M. Sigala, E. Christou, and U. Gretzel, *Social media in travel, tourism and hospitality: Theory, practice and cases*: Ashgate Publishing, Ltd., 2012.

[13] D. Leung, R. Law, H. Van Hoof, and D. Buhalis, "Social media in tourism and hospitality: A literature review," *Journal of Travel & Tourism Marketing,* vol. 30, pp. 3-22, 2013.

[14] A. Abbasi, T. H. Rashidi, M. Maghrebi, and S. T. Waller, "Utilising Location Based Social Media in Travel Survey Methods: bringing Twitter data into the play," in *Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, 2015, p. 1.

[15] M. Maghrebi, A. Abbasi, T. H. Rashidi, and S. T. Waller, "Complementing Travel Diary Surveys with Twitter Data: Application of Text Mining Techniques on Activity Location, Type and Time," in *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, 2015, pp. 208-213.

[16] J.-P. Onnela, S. Arbesman, M. C. González, A.-L. Barabási, and N. A. Christakis, "Geographic constraints on social network groups," *PLoS one,* vol. 6, p. e16939, 2011.

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

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

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

[20] T. H. Rashidi, Abbasi, A., Maghrebi, M., Hasan, S., Waller, T.S., "Exploring the Capacity of Social Media Data for Modelling Travel Behaviour: Opportunities and Challenges," *Transportation Research Part C: Emerging Technologies,* under review.

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

[22] H. Gao, J. Tang, X. Hu, and H. Liu, "Exploring temporal effects for location recommendation on location-based social networks," in *Proceedings of the 7th ACM conference on Recommender systems*, 2013, pp. 93-100.

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

[24] F. Girardin, F. Calabrese, F. D. Fiore, C. Ratti, and J. Blat, "Digital footprinting: Uncovering tourists with user-generated content," *Pervasive Computing, IEEE,* vol. 7, pp. 36-43, 2008.

[25] A. Majid, L. Chen, G. Chen, H. T. Mirza, I. Hussain, and J. Woodward, "A context-aware personalized travel recommendation system based on geotagged social media data mining," *International Journal of Geographical Information Science,* vol. 27, pp. 662-684, 2013.

[26] Y. Sun, H. Fan, M. Helbich, and A. Zipf, "Analyzing human activities through volunteered geographic information: Using Flickr to analyze spatial and temporal pattern of tourist accommodation," in *Progress in Location-Based Services*, ed: Springer, 2013, pp. 57-69.

[27] A. Pozdnoukhov 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, pp. 1-8.

[28] M. De Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu, "Automatic construction of travel itineraries using social breadcrumbs," in *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, 2010, pp. 35-44.

[29] T. Ichimura 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, pp. 110-115.

[30] A. Abbasi, Rashidi, T.H., Maghrebi, M., "How Universal Human Mobility Patterns Are?," *PloS One,* work in progress.

[31] K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg, "Chatter on the red: what hazards threat reveals about the social life of microblogged information," in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 2010, pp. 241-250.

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

[33] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring Millions of Footprints in Location Sharing Services," *ICWSM,* vol. 2011, pp. 81-88, 2011.