1-2016

On analyzing geotagged tweets for location-based patterns

Philips Kokoh PRASETYO
Singapore Management University, pprasetyo@smu.edu.sg

Palakorn ACHANANUPARP
Singapore Management University, palakorna@smu.edu.sg

Ee-peng LIM
Singapore Management University, eplim@smu.edu.sg

DOI: https://doi.org/10.1145/2833312.2849571

On Analyzing Geotagged Tweets for Location-Based Patterns

Philips Kokoh Prasetyo
Singapore Management University
80 Stamford Road
Singapore 178902
pprasetyo@smu.edu.sg

Palakorn Achananuparp
Singapore Management University
80 Stamford Road
Singapore 178902
palakorna@smu.edu.sg

Ee-Peng Lim
Singapore Management University
80 Stamford Road
Singapore 178902
eplim@smu.edu.sg

ABSTRACT

Geotagged social media is becoming highly popular as social media access is now made very easy through a wide range of mobile apps which automatically detect and augment social media posts with geo-locations. In this paper, we analyze two kinds of location-based patterns. The first is the association between location attributes and the locations of user tweets. The second is location association pattern which comprises a pair of locations that are co-visited by users. We demonstrate that through tracking the Twitter data of Singapore-based users, we are able to reveal association between users tweeting from school locations and the school type as well as the competitiveness of schools. We also discover location association patterns which involve schools and shopping malls. With these location-based patterns offering interesting insights about the visit behaviors of school and shopping mall users, we further develop an online visual application called Urbanatics to explore the location association patterns making use of both chord diagram and map visualization.

CCS Concepts

•Information systems → Mobile information processing systems;

Keywords

location-based patterns; urbanatics

1. INTRODUCTION

Motivation. Geotagged social media is becoming highly popular as many users today are equipped with using smart phones and are familiar with social media apps that run on these phones. For example, Foursquare and Instagram are very popular location-based social media where users perform check-ins and share photos taken at different locations respectively. Traditional social media platforms such as Facebook and Twitter also provide geotagging features for their users to share geotagged status updates and photos.

When geotagged social media content are made publicly available, they represent important data traces that can be mined for interesting patterns about human mobility and location association. One can in turn utilize these patterns to model the expected volume of human flows between locations, to evaluate locations for business opportunities, and to detect events that disrupt the normal patterns, We will elaborate some of these applications of analyzing geotagged social media content in Section 2.

Research Objectives. In this paper, we focus on analyzing the geotagged public Twitter data generated by more than 150,000 Singapore users. Our goal is to find associations between user tweeting locations and location attributes and associations between user tweeting locations for the purpose of understanding how users' tweeting locations in an urban city can tell us information about these locations. We demonstrate that location affects tweeting behavior of users. We also propose location association to reveal the location preferences of users who visit some common location. By considering the semantics of location, we even derive some interesting findings from location association patterns.

Although geo-coded Twitter data can be crawled using public Twitter APIs, selecting an appropriate subset of data to track over a meaningful time period and analyzing them are still challenging. The research on geo-coded social media is also at its infancy. Finally, in-depth location association analysis requires a good visualization tool.

In order to achieve our research goals, we thus make the following contributions:

- *Data crawling*: We have gathered a complete set of Twitter data from users located in Singapore during the month of July in 2012. With this complete set of Twitter data, we are able to find about 6000 users who are very active in posting geotagged tweets. This dataset allows us to extract location association insights from all geotagged tweets of these 6000 users.
- *Location association pattern mining*: We introduce the concept of location association pattern and its definition. We also show how location association pattern can be turned into location transitions which can be easily represented by a transition graph which can be visualized as a chord diagram.
- *Empirical analysis of location association patterns*: We conduct an empirical analysis of the location associa-

ICDCN '16, January 04-07, 2016, Singapore, Singapore

DOI: http://dx.doi.org/10.1145/2833312.2849571

tion patterns of users who post tweets in schools and in shopping malls. By considering the ranking of secondary schools in Singapore, we are able to show that: (a) Users from academically competitive schools are less likely to post tweets compared with users from less competitive schools; (b) When we divide the city of Singapore into five regions, users posting a tweet from some regions have high likelihood to post other tweets from the same regions; and (c) When we rank the secondary schools by academic competitiveness, users who post tweets in competitive schools are less likely to post tweets in shopping malls compared with users who post tweets in less competitive schools. These findings are important knowledge about behaviors of users tweeting from schools and shopping malls.

- *Visualization of location association patterns*: Finally, we have developed a visual analytics tool to visualize location association patterns at different locations or different types of locations. Furthermore, the tool supports interactive visualization useful in exploratory analysis of these patterns.

Paper outline. Section 2 covers a few important works on geotagged social media. Section 3 describes the data gathering step of our research and the analysis of tweeting locations. The dataset is used directly for deriving location association patterns in Section 4. Section 5 lists the interesting findings with supporting statistics and case examples. A brief description of a visual analytics tool to explore location associations is given in Section 6. Finally, we conclude the paper in Section 7.

2. RELATED WORK

Research on location information in Twitter data has been very active for a while. Many works study the location field of the user profiles. For example, Jurgens et. al developed a spatial label propagation technique to predict this location field. Unfortunately, Hecht et. al found that the location field can be easily tampered and as many as 34% users do not provide actual location information [8]. Graham, et. al also found the user profile location field does not always tally with the locations of geotagged tweets [7]. In this paper, we do not use the location field in user profiles. Instead, we employ the locations of geotagged tweets directly.

There are also other research on location information found in geotagged tweets. When a user actively generates geotagged tweets, we may obtain the location trajectory of the user. Such users are however very rare, and hence one has to aggregate the geotagged tweets from different users to derive some interesting insights. For example, Hong et. al discovered the content topics of Twitter based on tweet locations and language [9].

Beyond Twitter data that provides geotagged, there are many other works on other location-based social media platforms. For example, in [3], it has been shown that the locations of Foursquare like check-ins performed by a user can be affected by the user's home and work places, as well as his friendships. In [4], a study on trips made by short-term/long-term visitors versus that of local residents was conducted using the Foursquare's check-in data gathered from Singapore. It was shown that short-term visitors have distinctive check-in locations compared with long-term and local residents.

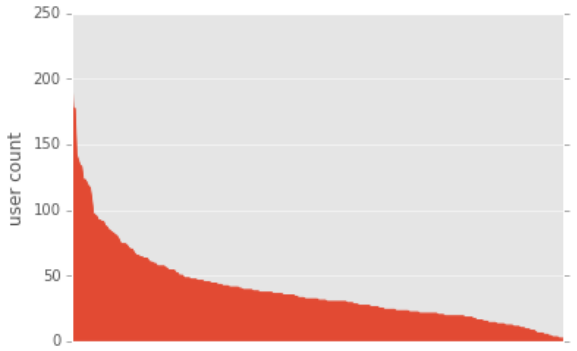

Figure 1: Distribution of user count over all schools

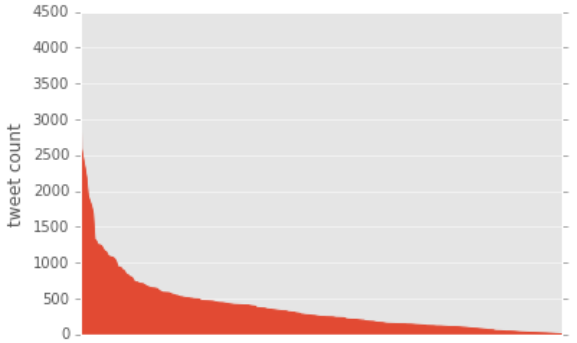

Figure 2: Distribution of tweet count over all schools

3. DATASET AND TWEETING LOCATION ANALYSIS

In this section, we first construct a Twitter dataset for conducting an empirical analysis of the user tweeting locations. Our goal is to derive location-based patterns linking tweeting behavior with some location attributes.

3.1 Dataset

In this study, we need a longitudinal Twitter dataset covering users within a geographical area. We select Singapore which is a city state with very active Twitter users. Starting from a small set of prominent Singapore seed users who are largely celebrity, news media and political Twitter accounts, we crawled their followers and followees who are also located in Singapore (as indicated in the user profiles). We repeated this snowball sampling for a few iterations until we are not getting more Singapore users. In total, we obtained about 150,000 Singapore-based public user accounts, who generate roughly 1.2 million tweets per day.

Utilizing Twitter API [1], we also collected nearly 950,000 geo-coded tweets from 38,646 users in July 2012. This suggests that only 2.6% of Singapore tweets are geo-coded and 25% Singapore users generate geo-coded tweets. We further selected active users with at least 30 geo-coded tweets in a month, resulting in 757,804 geo-coded tweets from 5,934 users who represent about 4% of all Singapore users we have.

We also collected geolocation and region information of schools and malls in Singapore based on the lists of Singa-

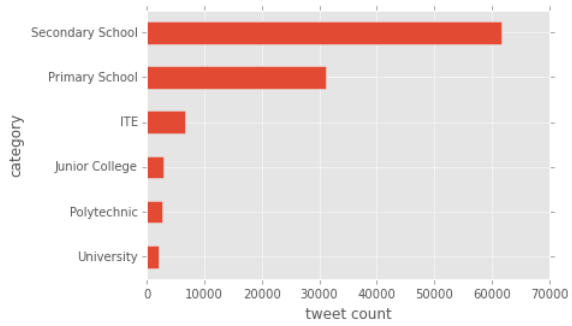

Figure 3: Distribution of tweet count over all school categories

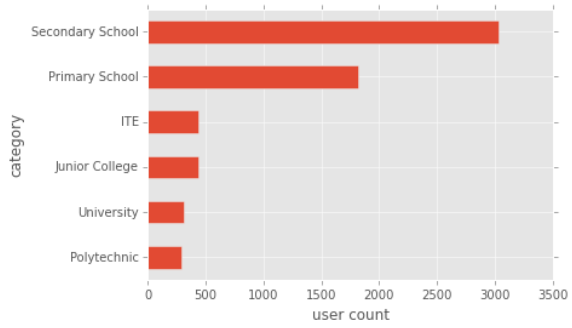

Figure 4: Distribution of user count over all school categories

pore schools[1] and malls[2] from Wikipedia. In total, we have collected 303 schools and 92 malls in Singapore. We also obtained the locations of these schools and malls which are used to determine location of geotagged tweets. As we only have a single location point for each school/mall, we empirically assign a geotagged tweet to the nearest school or mall within 200-meters radius from the tweet's coordinate.

We now present some distribution statistics of the geotagged tweets and their users. Figure 1 depicts the distribution of user count over all 303 schools in decreasing order. The figure clearly shows that most schools have very few tweeting users. The distribution of tweet count over schools (in decreasing order) in Figure 2 also shows that most schools have very few tweets. Due to space constraint, we do not show the user count and tweet count distributions of shopping malls.

3.2 Analysis of Tweeting Location Patterns

We now analyze the correlation between the tweeting volume and the type of schools. In Singapore, schools are categorized into primary schools (elementary schools), secondary schools, institutes of technical education (ITE), junior colleges, polytechnics and universities. ITEs are vocation training schools for students who could not gain admission into secondary schools. Junior colleges are schools that offer two years academic program prior to universities. Polytechnics are schools for graduates of secondary schools and junior colleges should they decide to be trained directly for industry jobs instead of further academic pursuits in universities.

[1]https://en.wikipedia.org/wiki/List_of_schools_in_Singapore
[2]https://en.wikipedia.org/wiki/Category:Shopping_malls_in_Singapore

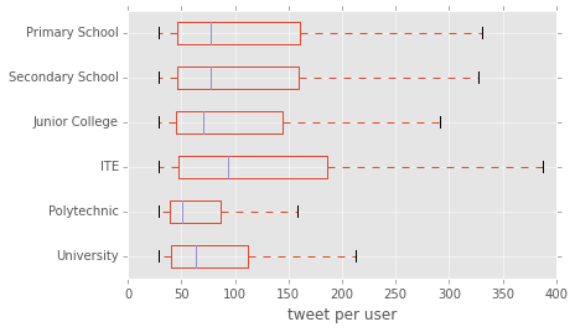

Figure 5: Distribution of tweets per user over all school categories

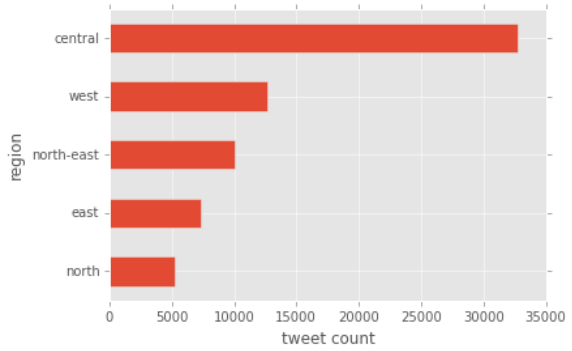

Figure 6: Distribution of tweet count over malls from different regions

The geotagged tweet count distribution in Figure 3 shows that most tweets are generated by users from secondary school and primary school categories. Much fewer tweets are generated from university, polytechnics, junior college and ITEs. These observations can be partially explained by the number of tweeting users from the different school categories as shown in Figure 4. Although primary schools should have more students than secondary schools, many of these students may not have been active on social media and may not own a smart phone. In 2012, about 21% of students are in ITEs, 46.5% in polytechnics, 27.7% in junior colleges, and 28.2% in universities [2]. From Figure 4, we infer that users from tertiary level education institutions which include polytechnics and universities have less active Twitter users.

As we drill down to the tweet per user statistics as shown in Figure 5, we further discover that even for the tweeting users from tertiary institutions, they generate fewer tweets than primary and secondary school users. The ITE users are most active compared to the rest with a median of about 80 geotagged tweets in a month.

In a similar manner, we analyze the volume of tweets and number of users tweeting from shopping malls of different regions, namely, central, east, west, north, and north-east. Not surprisingly, more tweets and users are found in many malls from the central region.

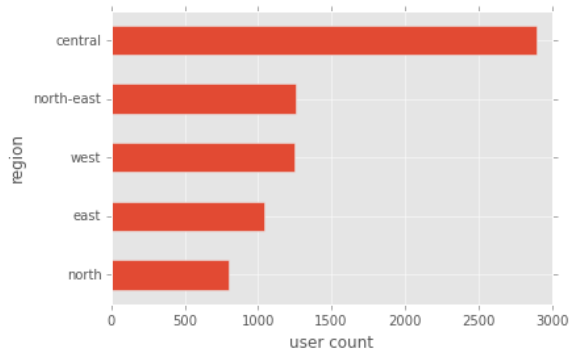

Figure 7: Distribution of user count over malls from different regions

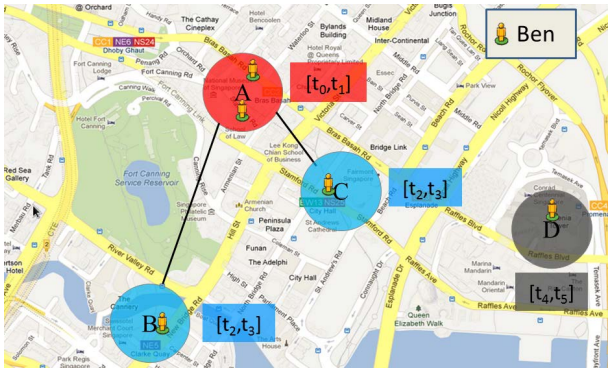

Figure 8: Associated locations analysis

4. ANALYSIS OF LOCATION ASSOCIATION PATTERNS

4.1 Location Association Pattern

Location association pattern represents the relation among locations. A **location association** x is a tuple $\langle l_r,\ l_t,\ [t_r^s, t_r^e],\ [t_t^s, t_t^e]\rangle$ where l_r and l_t represent the *reference location* and *target location* respectively, $[t_r^s, t_r^e]$ represents the *reference time interval*, and $[t_t^s, t_t^e]$ represents the *target time interval*. The support of a location association $x = \langle l_r,\ l_t,\ [t_r^s, t_r^e],\ [t_t^s, t_t^e]\rangle$ is defined by number of users who tweet at location l_r within the time interval $[t_r^s, t_r^e]$, as well as at location l_t within the time interval $[t_t^s, t_t^e]$.

Consider the example shown in Figure 8, user Ben tweets at location A during time interval $[t_0,t_1]$ on one day, B during the time interval $[t_2,t_3]$ on the same day and C during time interval $[t_2,t_3]$ on a different day. We say that both location associations $\langle A, B, [t_0, t_1], [t_2, t_3]\rangle$ and $\langle A, C, [t_0, t_1], [t_2, t_3]\rangle$ have support of 1 due to Ben.

When both reference and target time intervals are $[0000, 2359]$ (that is, time is no longer important), we simplify the location association tuple by dropping the time interval elements, i.e., $x = \langle l_r, l_t\rangle$. For the rest of this paper, we shall conduct our analysis without considering the reference and target time intervals. We shall conduct a time dependent analysis in the future work.

Now, consider a set of users in Table 1 each contributing a number of location associations. We generate from these data a *support matrix* by counting the number of associated location pairs as shown in Table 2. The row and column in support matrix represent reference and target locations respectively. Each cell value indicates the number of users supporting the corresponding location association.

Table 1: Toy example

User	Associated Location Instances
Amy	$\langle B, C\rangle$, $\langle C, A\rangle$, $\langle C, B\rangle$
Ben	$\langle A, A\rangle$, $\langle A, B\rangle$, $\langle A, C\rangle$
Chris	$\langle B, C\rangle$, $\langle B, D\rangle$, $\langle D, B\rangle$

Table 2: Support matrix

Reference Location	Target Location			
	A	**B**	**C**	**D**
A	1	1	1	0
B	0	0	2	1
C	1	1	0	0
D	0	1	0	0

Table 3: Transition probability matrix

	A	**B**	**C**	**D**
A	0.33	0.33	0.33	0
B	0	0	0.67	0.33
C	0.5	0.5	0	0
D	0	1	0	0

Once the support matrix is generated, we perform normalization to derive transition probabilities. Table 3 shows the normalized transition probability matrix. The normalization is performed on every reference location (row basis).

4.2 Chord Diagram for Visualizing Location Association Patterns

From associated location instances, we can derive activity patterns and visualize them in a circular layout called chord diagram [10]. This diagram provides rich summary of probability transition data, and thus we adapt it for exploring relationships between locations.

Chord diagram presents the data in circular layout. The chord segments in the diagram represents locations or location types, and are assigned unique colors. The ribbon connecting two segments represents bi-directional correlation between two locations or location types. Each ribbon between segments X and Y represents the transition from X to Y and that from Y to X. The thickness of the X's end of ribbon represents the probability value of transition from X to Y, while that of the Y's end of the ribbon represents the probability value of the transition from Y to X. The color of the ribbons follows the color of the location (or location type) with higher outgoing transition probability.

Figure 9 depicts a location type chord diagram where each chord segment represents a group of malls or schools in one of the five geographical regions. D3.js library is used to draw chord diagram. We shall elaborate the findings in this figure in Section 5.1.

5. EMPIRICAL FINDINGS

Next, we present some interesting findings about the location association patterns discovered from our Twitter dataset.

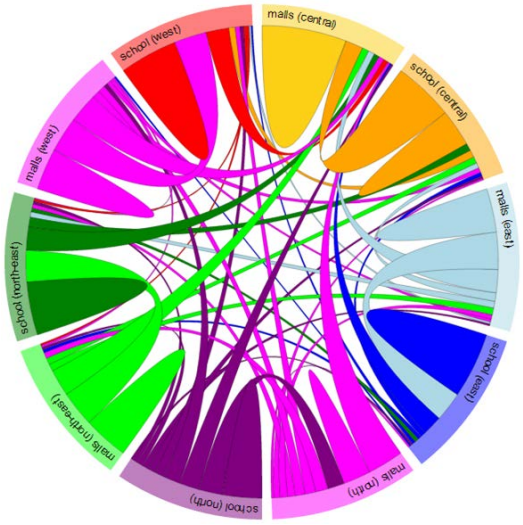

Figure 9: Chord diagram of mall-in-region and school-in-region segments

We aim to find neighborhood effect and school ranking effect on the locations co-tweeted by users.

5.1 Neighborhood Effect

Research has found most users are much more likely to visit places near their homes and work places [5]. Gonzalez et al.[6] also found user mobility follows reproducible patterns with high degree of temporal and spatial regularities.

In this study, we want to evaluate how the users tweeting from schools are related to those tweeting shopping malls when these schools and malls are from different geographical regions. As mentioned earlier, we divide Singapore into Central, East, West, North, and North-East regions. From our Twitter dataset, we first extract the location associations of all users. These location associations are turned into location-type-region associations before we derive the transition probabilities between location-type-regions (e.g., mall-in-central, school-in-northeast, etc.).

We first examine the self transition probabilities of malls in Figure 9. As the outgoing transitions of each chord segment are ordered by increasing outgoing transition probability from left to right (with reference to the center of chord diagram), the rightmost transition of a chord segment is one with the highest probability. All mall-in-region segments except the mall-in-north segment have most of their tweeting users tweeting in other malls in the same region. The malls in the north region tend to see most of their tweeting users at the malls in the central region. The figure also shows that malls in the central region have strong attraction to users from both malls and schools of other regions.

School wise, it is interesting to find the school-in-central segment having the largest transition probability going to the mall-in-central segment. This suggests that many users tweeting in schools of the central region show up in the mall of the same region. Whether this pattern may lead to weaker academic performance is a topic for our future research.

5.2 School Ranking Effect

Based on the admission criteria of 131 secondary schools in Singapore, we now divide the secondary schools into five

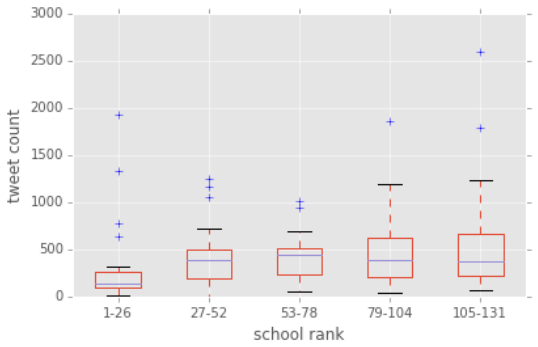

Figure 10: Tweet count for school rank bins

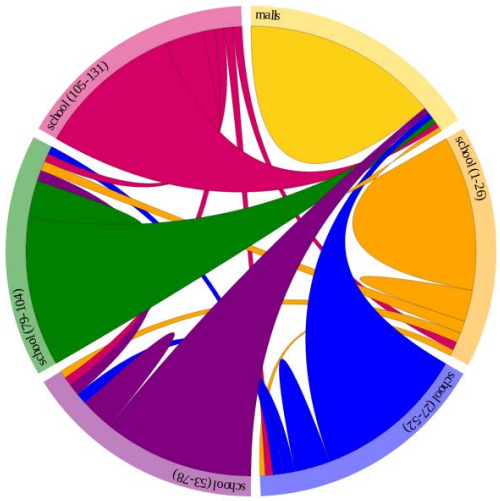

Figure 11: Chord diagram (school rank and malls)

school rank bins. The admission of these schools are determined by a nation wide examination in the final year of primary school education. The higher cut-off for the examination marks, the higher (or smaller) rank the school. The five school rank bins cover the rank 1 to 26 schools, rank 27 to 52 schools, rank 53 to 78 schools, rank 79 to 104 schools, and rank 105 to 131 schools respectively. Again, we derive the transition probabilities between school rank bins and a mall segment (which represents all malls) from location associations between schools and malls.

The tweet counts of rank 1-26 schools are considerably fewer than those of lower ranks as shown in Figure 10. Figure 11 depicts the result transition probabilities. The figure shows that other than the top school rank bin (covering the most competitive schools), the other school rank bins have largest outgoing transition probabilities to the mall segment. This suggests that users from the competitive schools tend to visit the malls less. In contrast, users from the weaker schools are far more likely to visit the malls.

6. VISUAL ANALYTICS OF LOCATION ASSOCIATION PATTERNS

To aid the exploration of location association patterns, we have developed a visual analytics tool known as Urbanatics

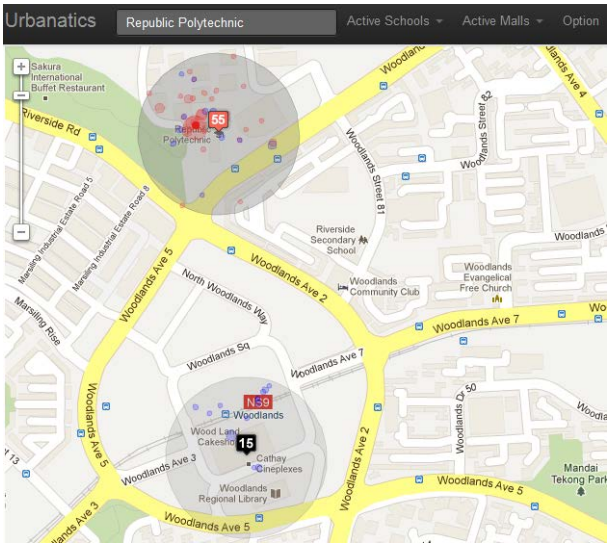

Figure 12: Map visualization

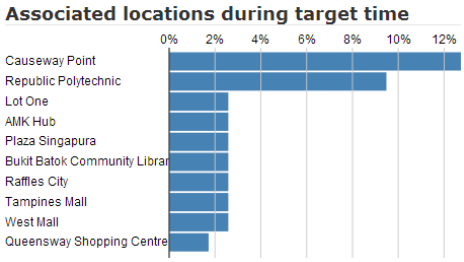

Figure 13: Most popular associated locations during target time interval

to allow users to interactively explore associated locations on the map. Users have three alternatives to select reference location: (1) use search box in the toolbar to search reference location. (2) click a random location on the map to get all available reference locations near the selected locations. (3) select from recommended active schools and malls. Users can click option button to open a panel to select reference and target time interval.

6.1 Visualization of Associated Locations

Figure 12 shows associated locations in Urbanatics. The associated locations visualization is shown directly on map using Google Maps API. Area covered by red marker denotes the reference location. Area covered by black markers denotes the target locations. Number on the markers shows the number of users in the area. Circles represent centroid of user locations in the area. Red color indicates that the users tweet on the reference time interval, while blue color indicates that the users tweets on the target time interval. The size of the circle represents user's tweet activities.

Beside the interactive visualization on map, Urbanatics also provides summaries of selected associated locations such as associated locations distribution, and most popular associated locations during target time. Figure 13 shows the example of top 10 most popular associated locations during target time interval.

7. CONCLUSIONS

Sensing and learning the urban behavioral patterns of users through online social media is an emerging but challenging research topic. In this paper, we study such behavior using location-based patterns found in geotagged social media. By associating location type and locations of user tweets, we found users from institutes of technical education are the most active by number of tweets per user. In contrast, users from polytechnics and universities are less active both by number of tweets per user and proportion of tweeting users.

We also introduce location association pattern and use that to study relationships between locations and location types. We are able to find malls in the central region of Singapore attracting many users from malls and schools of other regions. Malls from the north region particularly see more their tweeting users visiting malls in the central region. Using location association patterns, we also find the relatively smaller proportion of users from competitive secondary schools visiting shopping malls compared with the less competitive schools. Finally, we feature the Urbanatics visual analytics tool that visualizes the location association information using map based interface and chord diagram.

For future work, we plan to extend the empirical study to include time dependent location association patterns. The study can include a comparison of weekday and weekend patterns, as well as comparison of patterns with different reference and target time intervals. We may also extend the study to include locations beyond schools and malls.

Acknowledgement

This work is supported by the National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

8. REFERENCES

[1] Twitter API. https://dev.twitter.com/.
[2] Singapore's ministry for education, education statistics digest. http://www.moe.gov.sg/education/education-statistics-digest/files/esd-2014.pdf, 2014.
[3] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *KDD*, 2011.
[4] W.-H. Chong, B. T. Dai, and E.-P. Lim. Not all trips are equal: Analyzing foursquare check-ins of trips and city visitors. In *ACM COSN*, 2015.
[5] N. Eagle and A. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7), 2009.
[6] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
[7] M. Graham, S. A. Hale, and D. Gaffney. Where in the world are you? geolocation and language identification in twitter. *The Professional Geographer*, 66(4):568–578, 2014.
[8] B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *CHI*, 2011.
[9] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *WWW*, 2012.
[10] M. I. Krzywinski, J. E. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: An information aesthetic for comparative genomics. *Genome Research*, 2009.