Towards Situational Pattern Mining from Microblogging Activity

Nathan Gnanasambandam
Xerox Innovation Group
Xerox Corporation
800 Phillips Road
Webster, NY 14580
nathang@xerox.com

Keith Thompson
Watson Institute of Systems Excellence
Binghamton University
Binghamton, NY 13902
kthomps6@binghamton.edu

Ion Florie Ho
Watson Institute of Systems Excellence
Binghamton University
Binghamton, NY 13902
iho1@binghamton.edu

ABSTRACT

Many useful patterns can be derived from analyzing microblogging behavior at different scales (individual and social group). In this paper, we derive patterns relating to spatio-temporal traffic flow, visit regularity, content and social ties as they relate to an individual's activities in an urban environment (e.g., New York City). We also demonstrate, through an example, methods for reasoning about the activities, locations and group structures that may underlie the microblogging messages in the aforementioned context of mining situation patterns. These individual and group situational patterns may be very crucial when planning for disruptions and organized response.

Categories and Subject Descriptors

H.2.8. [**Database Applications**]: Data mining; K.4.2 [**Computing Milieux**]: Computers and Society – *social issues*; H.4 [**Information Systems Applications**]: Miscellaneous.

General Terms

Experimentation, Measurement.

Keywords

Social media mining, Information management, situational awareness

1. INTRODUCTION

Social microblogging activity (on such sites as Twitter, Tumblr etc.) has been used for various research purposes including detecting emerging topics of interest in real time [1] and predicting the success of movies at the box office [2]. Microblogging data also contains valuable information that māy be analyzed and applied toward situation pattern discovery. For instance, entity extraction and location estimation from microblogging activity have been used to detect earthquakes in real-time and, thus, alert people to possible danger [3]. In this paper, we analyze a trace from a popular microblogging platform and attempt to deduce the characteristics of human movement in an urban environment. Furthermore, once we are situationally aware of inter-block human flow and regularity patterns in a city (we consider New York City) and other contextual patterns based on location, we may be able to relate it to how it will affect an individual's activities in the event of a disruption or disaster.

2. MINING MIROBLOGGING FEEDS

Over a period of about seven months in 2010, microblogging feeds generated by a community of users in the greater New York City area (within the rectangular grid bounded by {40.45° latitude, -74.5° longitude} at the southwest corner and {41° latitude, -73.5° longitude} at the northeast corner) were collected. These feeds contain spatio-temporal information, conversational content and social ties between individuals. Our first objective is to isolate movement patterns in specific areas of the city. From these patterns, we hope to understand likely movement patterns of people or social groups. Our second objective is to derive a likely home-base (i.e. the centroid of the most frequented location cluster) at an individual level and a characterization of various activities that this individual may be involved in either alone or with a group. With the help of the home-base we determine which grids are regular with respect to a given set of users. Subsequently, we also utilize the content in the microblogging activity to inform us further about the situation. Our hypothesis is that we can utilize the knowledge so mined to reason about an individual's situational risk at a particular urban location.

3. SITUATIONAL PATTERN MINING

3.1 Data Description

As mentioned above, we consider a trace of microblogging activity in the greater New York City area. The data consists of GPS coordinates from which users either microblog or "check in". From the data, it is apparent that the people are moving from place to place, often as a group. For the purpose of this study, users who have posted for a hundred times or more in the selected area have been retained, along with the GPS data, content and time stamp for each post. The resulting subset contains 1494 users. A typical post is in the format of:

> Chilling in <event A> suite - thx guys! (@ Madison
> Square Garden w/ <person B>) <hyperlink C>

The entities of interest in this post include the event '<event A>', the place 'Madison Square Garden', the friend '<person B>' and the attached hyperlink. From the messages, over time, we can infer the social ties between people. The attached hyperlink, '<hyperlink C>' provides further information about the particular place in the post.

3.2 Flow Patterns

The geographic area being studied covers over 5000 square kilometers. In order to establish the flow patterns at a granular level we divide the map area into 5500 virtual grids, each with an area of about one square kilometer. Any situational pattern we identify is with respect to a single grid.

We first characterize what we mean by a flow. We delineate all time-stamped and GPS marked activity at the granularity of a day or portions thereof. During the course of a day, we track the movement of people from place to place, or alternately from grid to grid using pairs of messages (m_i, m_{i+1}) of the same person taken from the trace.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

These daily movement patterns form the basis for computing the in- and out-flows with respect to every grid. For each grid we compute the number of messages that either terminate at or emanate from it. We also keep track of the cardinal and intermediate directions from which these movements originate or proceed towards. As a result, we get two 8-vectors Fin_j and $Fout_j$ constituting the in- and out-flows for each grid j. In Figure 1, we show the flows computed for a certain portion of New York City.

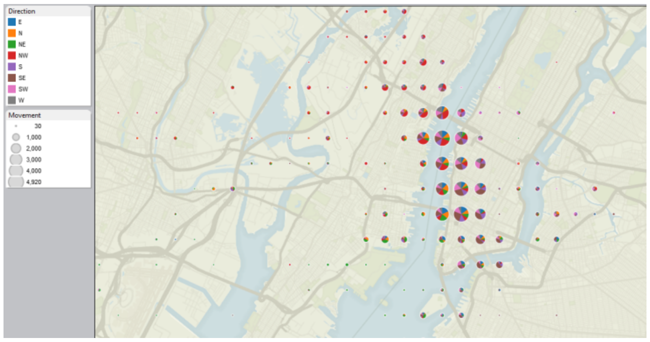

Figure 1. Inflows on a portion of New York City (size is proportional to number of microblogging messages)

From Fin_j and $Fout_j$, we know the proportion of flows in different directions for every grid. We represent these fractions as a pie-chart at the center of each grid. Not all grid locations are shown as they do not exceed the minimum threshold of messages (in this case at least 30 visits by our population of 1494 subjects). These flow numbers give a precise indication of where people come from or go to. For example, in Figure 2 we show two inflows – the one of the left shows inflows in every direction whereas the flow from the right grid indicates that people travel from the East, South-East, South, South-West and West directions. These pie-charts also show the change in flow conditions from the previous observed month. While many factors including where people reside influence these flows, we also know that the grid indicated in Figure 2(a) is close to a many dining spots and likely attracts people from various directions. Using just microblogging activity, we have also been able to compute the changes of incoming traffic patterns relative to the previous month from the various directions.

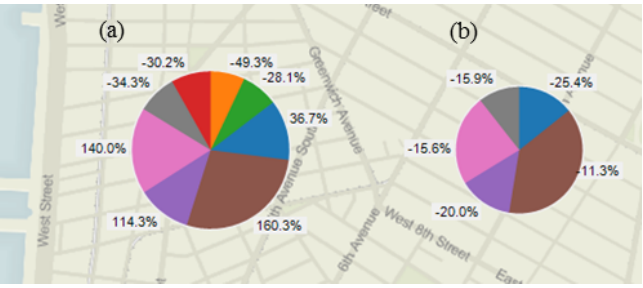

Figure 2. (a) (clockwise from orange) Fraction of inflows in different directions {N, NE, E, SE, S, SW, W, NW} and (b) (clockwise from blue) Fraction of inflows in the directions {E, SE, S, SW, W} and changes from the previous month

Furthermore, from just the microblogging activity we can also compute the net flows (i.e. $F_j = Fin_j - Fout_j$) with respect to the same grids as shown in Figure 3. With F_j, we can estimate where people are likely to arrive or depart. Regions with positive netflow (as indicated by the color blue in Figure 3b) indicate more people are likely to aggregate (and microblog) from those locations. To the contrary regions with negative netflow may be closer to transportation routes. With such flow information, we can reason about a specific group of individuals. Such information will not only inform us to the likely change in movement behavior but also which grids specific groups of people visit or leave. This also raises an interesting question as to why people stay close to certain grids. In the next section, we attempt to infer some aspects that further characterize these grids.

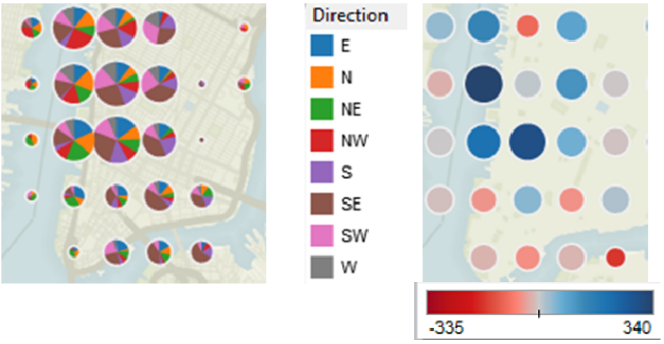

Figure 3. Net flow information for a portion of Manhattan, NY (a) indicates the flow over a busy portion of Manhattan (b) inflow-outflow for the same grid for the same region within Manhattan

3.3 Regularity Patterns

With respect to the same grid structure, we can determine if people visit some grids more regularly than others. The grids that are regularly visited by a group of users are also called non-transient grids. These grids might be areas that are closer to a person's home or workplace. By the same token, grids that are primarily used as "pass-through" grids are called transient grids.

In [4], we propose an algorithm to differentiate between transient and non-transient grids. An example of such regularity behavior is shown in in Figure 4. Determining regularity is predicated on what the likely operating base of a user is and, as shown in [4], estimates based on microblogging activity are possible.

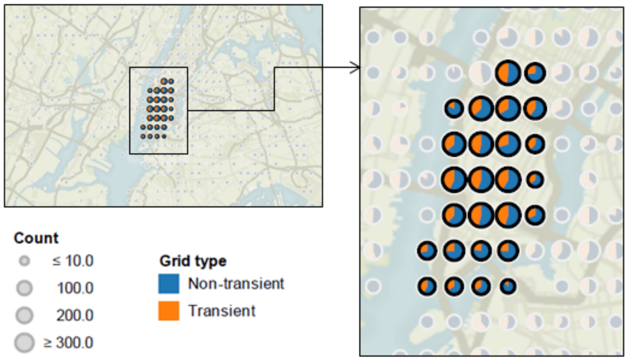

Figure 4. Transient and Non-transient grids for a group of people visiting Manhattan, NY

At the core of estimating whether some grids are transient or regular with respect to visits by a person or group is the concept of a home-base. In our case, the home-base is estimated by performing density-based clustering (using DBSCAN) over all coordinates visited by a user to determine the most probably location(s).

3.4 Content and Social Patterns

Even microblogging may have enough content to establish patterns on the basis of what the users do along with their social circle. In this particular case, we have been observing what areas were visited by our target population. For example, the user A has been frequenting a lot of restaurants and bars in the New York City area. Shown in Figure 5 is a tag cloud of the various categories of content mentioned in the microblogging activity representing the relative importance of various aspects, mainly restaurants for the two users A and B.

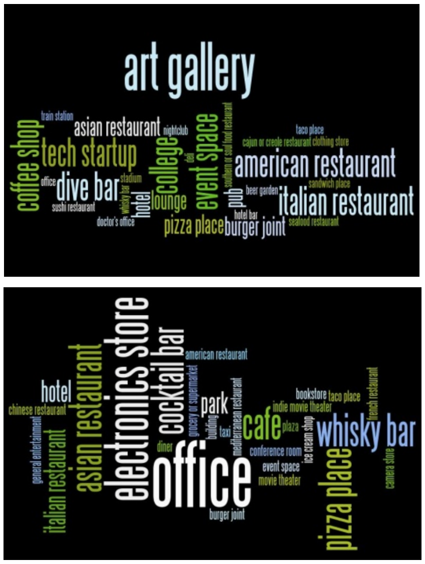

Figure 5. Tag cloud of user A (above) and user B (below)

User B is in the social circle of user A and has many similar dining interests as mined from the relative frequency of content included in microblogging activity. One difference that stands out is the interest of user A in tech startups and art, whereas User B tends to talk about "office" more. While these characterizations are preliminary, it can be observed that interest patterns on topics such as dining, work and art can be established relative to each other.

In addition to inferring information about the categories of places visited from the actual content of microblogs, we are also able to associate the various categories with both grid locations and social networks . For example, Figure 6 shows how location, content, and friendship are related to each other for user A.

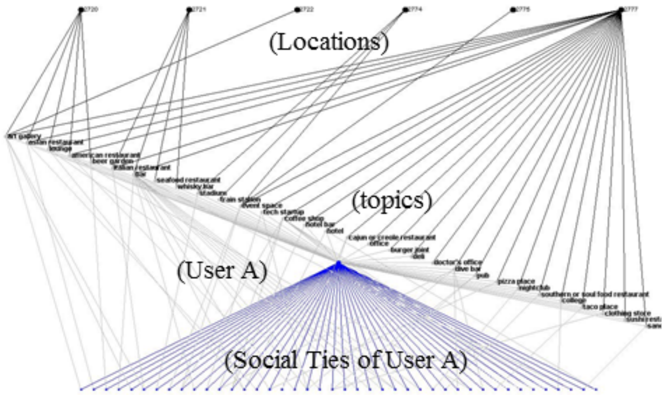

Figure 6. Connecting content (dining situations), grids (locations) and social circles

Examples of activities indicated in Figure 6 include {American Restaurant, Dive Bar, Wine Bar, Snack Place, Coffee Shop, Mexican Restaurant}, {Office}, {House}, {Nightclub}. Furthermore, the social activities (in this case primarily dining) of people as they move around in groups can also be determined. This is shown in Figure 7.

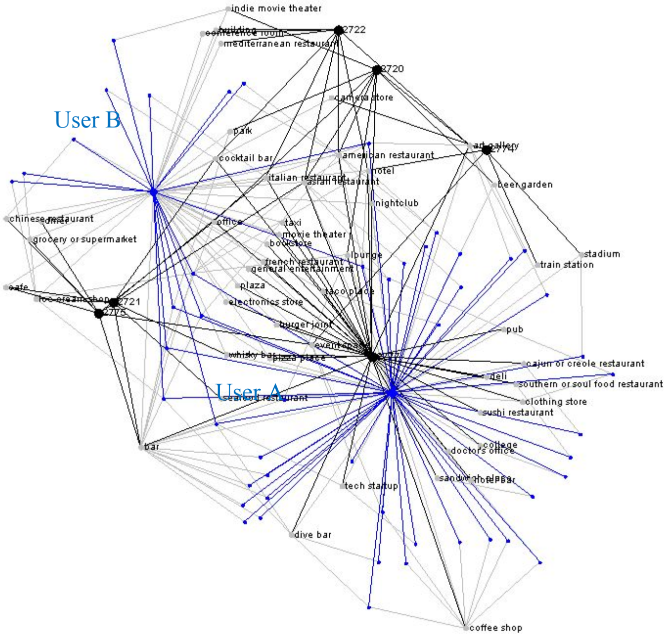

Figure 7. Content and social activities of Users A and B

4. SITUATIONAL PATTERNS

Our analysis has been able to unearth many situations from just mining microblogging activity. Combined with knowledge about other occurrences (such as traffic conditions, disruptions or disasters), the mined situational patterns whether personal, social or spatio-temporal can help with any sort of response. We discuss an example below extracted for the month of September 2010 for user A and the group this user belongs to.

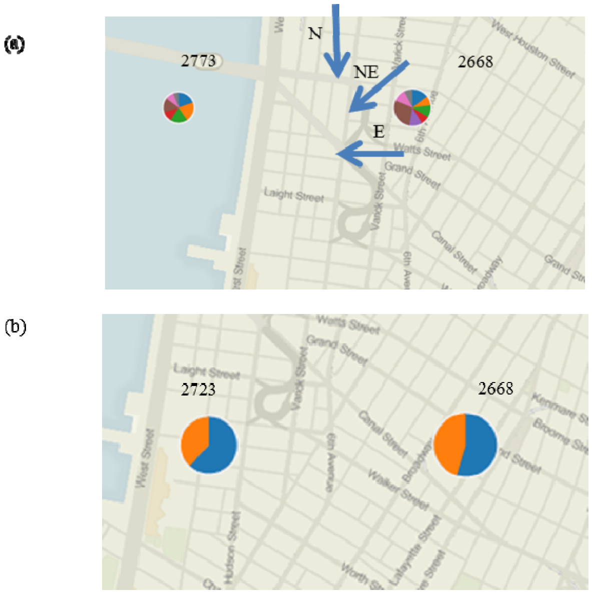

Figure 8. Flow and regularity patterns as it relates to Users A and B

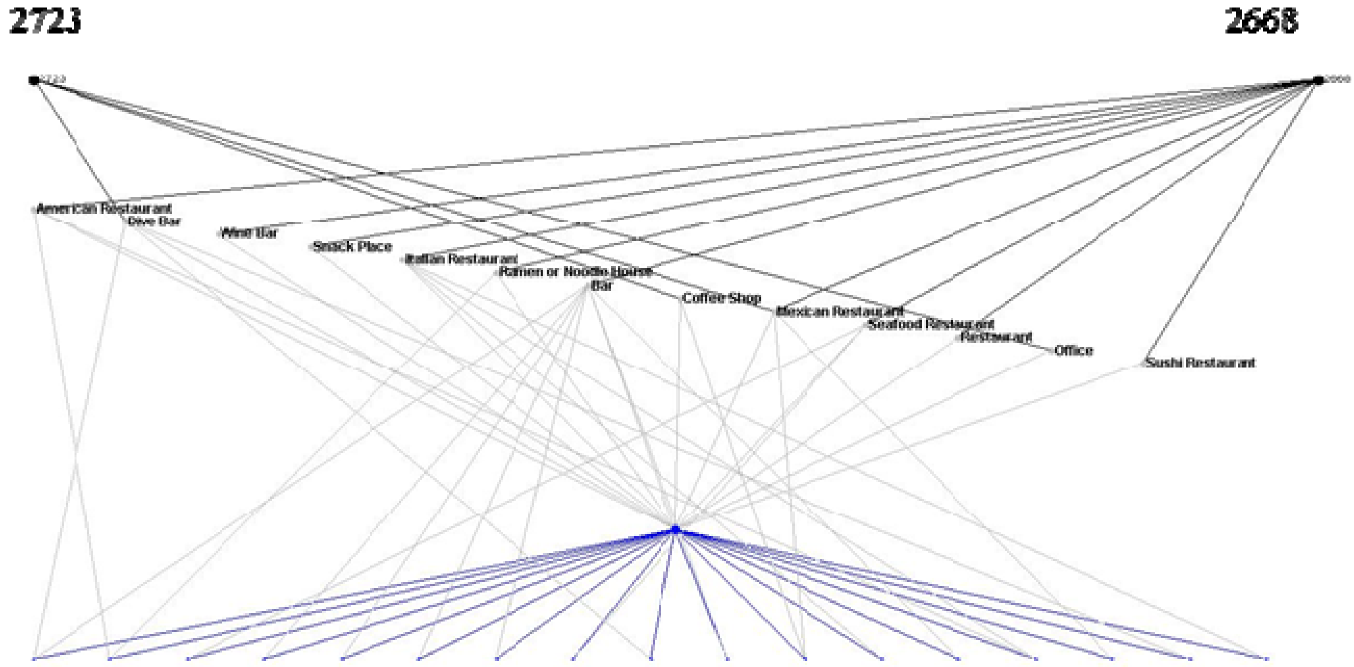

Figure 9. Content Pattern for User A

From the situational patterns (flow, regularity and content) we can surmise that for user A, grids 2773 and 2668 feature prominently in the activities. Firstly it should be noted that grid 2773 has many restaurants and grid 2668 has a number of transit points (i.e. subway routes and stations), office buildings and schools. We estimate that it is highly likely that grid 2773 acts as a frequent dining destination for user A while either of grid 2773 or 2668 is a work location. Our estimate is supported by the following individual and group situational patterns:

a. From the flow patterns outlined in Figure 8a, it is apparent that people were traveling mostly from {N, NE, E} directions into grid 2723. This indicates that user A may be traveling from 2668 to 2723 for the same reason that most others are, i.e. for dining.

b. The regularity pattern further supports the point about both grids 2773 and 2668 being very important for user A. We know that for user A, either of 2668 and 2723 is a home-base/operating base as indicated by the clustering algorithm DBSCAN. We also know that for the overall group these two grids are non-transient (see Figure 8b), again indicating that this grid pair are regular locations.

c. Both aforementioned reasons are reinforced by the content patterns shown in Figure 9. For the month of September 2010, the two grids have the highest incidence of dining locations and some mentions of the activity "office". This again points to these two locations being preferred work and dine-out destinations.

d. For user B who is the social network of user A, grid 2668 is a home-base. Additionally user A and user B are co-mentioned in many messages/activities and share many friends in common (as shown in Figure 7).

While automating the aforementioned reasoning is part of ongoing work, we have established some techniques to mine microblogging activities. Based on the patterns unearthed we can establish the likelihood of users and their social groups being involved in a subset of activities, locations and the flow directions of their movement. Further we can compare these individual or group likelihoods with the bigger population. This kind of analysis, we believe, can be utilized in locating and determining response strategies in situations that may involve disruptions. Furthermore, these response strategies can be targeted to the individual or their social group.

5. CONCLUSIONS

In this paper, we have established techniques to mine social microblogging activity from the perspective of flow, regularity (transience) and content. We have analyzed data (essentially millions of microblogging messages) from an urban environment. We also established a chain of reasoning that may assist with unearthing where and what a user may be involved in and with whom. These individual and group situational patterns may be very crucial when planning for disruptions and organized response. In particular, it may provide the capability to generate personalized evacuation planning and disaster recovery strategies. Using this work as the basis, future work will cover disaster planning aspects and how plans may relate to individuals tactically and strategically.

6. ADDITIONAL AUTHORS

Additional Authors: Sarah S. Lam (Binghamton University, email: sarahlam@binghamton.edu) and Sang Won Yoon (Binghamton University, email: yoons@binghamton.edu).

7. REFERENCES

[1] Cataldi, M., Di Caro, L., and Schifanella, C. 2010. Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation. *Proceedings of the Tenth International Workshop on Data Mining* (Washington, DC, July 25), 1-10.

[2] Asur, S., and Huberman, B.A. ., 2010. Predicting the future with social media. In *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology* (Toronto, Ontario, August 31 – September 3), 492-499.

[3] Sakaki, T., Okazaki, M., and Matsuo, Y. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on the World Wide Web* (Raleigh, North Carolina, April 26 – 30, 2010). WWW'10. ACM, New York, NY, 851-860. DOI= http://doi.acm.org/10.1145/1772690.1772777

[4] Ho, I.F., Gnanasambandam, N., Thompson, K., Lam, S., and Yoon, S.W. 2012. Mining transient vs. non-transient user movement behavior in urban environments. In *Proceedings of the Industrial and Systems Engineering Research Conference* (Orlando, FL, May 19 – 23). ISERC'12.