Ground-truthing spatial activities through online social networking data

G. D. McKenzie* and M. Raubal**

*Department of Geography, University of California Santa Barbara 1832 Ellison Hall, Santa Barbara, CA, USA
Email: grant.mckenzie@geog.ucsb.edu
**Institute of Cartography and Geoinformation, ETH Zürich, Switzerland
Email: mraubal@ethz.ch

1. Introduction

Recently, there has been considerable growth in the area of online social networking (OSN). In just six years, the social application *Facebook* has amassed over 900 million active users communicating in over 70 languages (Facebook Statistics 2012). Earlier this year, the location-based social network *Foursquare* reported over 3 million *check-ins* (Crowley 2012). Advances in mobile technology have integrated OSN even deeper into our everyday lives. Users are no longer tethered to their desktop computers in order to interact with their extended social network. The new world of mobile, ubiquitous computing allows them to instantly communicate with their friends through a *check-in* at the bus stop, *photo* from a football game, or *link* to the latest tracks from their favorite musicians.

Increasingly, social interactions are occurring online. Past, present, and future activities are often broadcast through OSN applications as a means of large-scale social information transfer and with the intention of increasing one's social worth (Ellison et al. 2007, Portes 1998). But how reliable is this information? From a geospatial perspective, what types of location information are being presented and at what resolution? And importantly, how accurate is the mentioned activity location in terms of real-world activities?

This extended abstract presents the initial stages of ongoing study designed to answer these questions. Preliminary results indicated that real-world activity locations play a substantial role in online social network contributions and while these data may not be 100% accurate in predicting real-world activities, it is expected that certain biographical and social factors can be matched to increase accuracy. This research offers a coarse method for measuring the truthfulness of online social network updates as they relate to user activities and real-world locations.

2. Related Work

Recently, online social networks have become a hot topic in the fields of social as well as (geo)spatial research. Not surprisingly, significant findings have come from studies in both private (social networking companies) and academic fields. Much of the geo-social related research has focused on the concept of "geo-tagging" or "checking-in" to a location. A number of studies have explored the usefulness of social network data in predicting future activities (Chang et al. 2011, Cheng et al. 2011) while others have investigated the effectiveness of travel trajectories in determining a user's hometown or place of residence (Liben-Nowell et al. 2005, Backstrom et al. 2011). Research in the transportation field has examined the strength of social ties in understanding travel behavior (Carrasco & Miller 2006) and inversely, a greater understanding of movement patterns and spatial behavior have proven to give insight into social intentions (Kiefer et al. 2010).

While most of these studies involve the examination and use of online communication, surprisingly few have ground-truthed the data with real-world travel or activity data. While it may be possible to predict an individual's location based on her previous *check-in* history, is this individual actually at the predicted location? And for that matter, was this person at any of the *check-ins* previously reported by the social networking application?

A number of previous studies have explored the idea of activity-based ground-truthing given real-world social survey information (Ahas 2005), but to the authors' knowledge, very little, if any research has investigated the position of online social networking data in determining an individual's real-world location.

3. Preliminary Methods

As a first step to exploring the correlation between online social network contributions and real-world activities, a preliminary study was conducted. 30 participants (15 female) between the ages of 20 and 45 installed the *Social Prediction* Facebook application and completed a three-week self-reported activity diary. Participants for this study were recruited through a snowball sampling method where acquaintances of the principle researchers recruited participants from their acquaintances and so on.

The purpose of the *Social Prediction* Facebook application was to allow researchers access to all information posted on a participant's Facebook *Wall.* Participants first granted access to this application and were informed that researchers would only have access to their basic profile information (age, gender, education, hometown, etc.) as well as data posted on their *Wall*[1] (typically only visible to friends). OSN information was gathered through this application for a period of three weeks before being removed from the participants' accounts.

Along with this application, participants completed a three-week self-reported activity diary. The diary consisted of an online calendar (similar to *Google Calendar*) on which participants were asked to provide an hourly account of their day. Each entry consisted of a start time, end time, location, and activity. For the purposes of this study, participants were informed that an activity was defined as "a change in location."

4. Preliminary Analysis

Of the 737 text-based *updates*[2] contributed by 28 participants (data from 2 participants was found to be corrupt), 94 contained information related to activity locations. The activity related updates were extracted using qualitative analysis performed by the principle researcher. Natural language processing was attempted (using NLTK 2.0[3]), but abandoned due to less than satisfactory results.

The 94 activity related updates were sorted and categorized based on whether the updated was in reference to a past, present, or future activity. Additionally, the updates were grouped by spatial resolution. The groups included: *Specific Building, City, In Transit, Limited High Frequency (non-specific "bar", "restaurant", etc.) and Limited Low Frequency (non-specific "school", "hospital", etc.)*. Lastly, categories were designated for

[1] Private messages, photo albums, and 3rd party applications were excluded.
[2] For the purposes of this study, "updates" include *status updates, photo captions, link captions, video captions,* and *checkin captions* (note that the *checkin* location was not used, only the text associated with the *checkin*).
[3] Python Natural Language Toolkit (nltk.org)

temporal forecasting (how far in the past or future did/will the activity occur?). These categories included: *Day, Week, Weeks, Month*, and *Unknown*.

5. Preliminary Results

The first outcome of this research is that online social network users often mention real-world activities online (this preliminary study found that 13% of updates contained some evidence signifying a real-world activity). Furthermore, 90% of these *mentioned activities* are true to real-world activities (based on participant's self-reported activity diaries). It is important to note that given that the period of study was only three weeks, any temporal forecast of a month or greater could not be confirmed through the participant's activity survey and was therefore not included in the above measurement of truthfulness.

Provided the number of OSN updates mentioning real-world activities, and the total number of OSN updates per person, an *OSN activity ratio* was assigned to each participant given as the number of *real-world mentions* divided by the total number of *OSN updates*. A linear regression model was then used to explore the influence of profile data (age, gender, post frequency, number of friends, and frequency of checkins) on this *OSN activity ratio*. Additionally, a crosstab / correlation matrix was generated to assess the correlation between variables.

Unfortunately, the correlation between the explanatory variables was quite low leading to a poor estimation of the linear regression model (r^2 = .026) and no significant contributing variables. The small sample size is most likely to blame for the lack of significance in the profile data contributing to the *OSN activity ratio* results. While this preliminary sample provided inconclusive results, a larger sample size would allow for an increase in the number of explanatory variables to be included in the model as well as more definitive results.

7. Future Analysis

Given the above sample limitations and inconclusive results, the next step in this research involved reversing the research question. Instead of asking *how many online updates were related to real-world activities*? we ask, *what types of real-world activities resulted in an online update?* Rather than focusing on the small sample size of 28 participants, the 3200 individual activities and 94 related online updates became the focus of analysis.

The 3200 individual activities were categorized using qualitative methods (3 individuals manually categorized each activity) into ten categories: *work, rest, self-maintenance, shopping, eating, drinking, fitness, errands, entertainment, other home, and other non-home.* While this research is still ongoing, preliminary results show that the activities categorized as *eating* or *entertainment* significantly contributed (*P* values of 0.029 & 0.007 respectively) to the likelihood of an online update.

8. Future Work & Limitations

Future work in this area includes a larger (number of participants) and longer (in duration) study. The self-reported activity survey should also be enhanced to include some level of participant tracking (e.g., GPS enabled mobile phones). A wider range of social networking applications will provide more breadth to the study and offer the ability to generalize results on a larger scale.

This type of research does not come without limitations and concerns. Privacy and data access are two major concerns when dealing with these types of data (Barkuus & Dey 2003, Bulgurcu et al. 2010). Most social networking data available online are subject to strict privacy guidelines and are often inaccessible to the general public. Additionally, it must be understood that information contributed to online social networks is inherently biased, volatile, and ambiguous. Humans can be indecisive and untrustworthy, and online contributions reflect this (one purpose of this research). Given the current limitations of natural language processing algorithms, this research required manual processing and entity extraction. While this is considered an acceptable qualitative process in this area of research it is far from ideal in terms of efficiency.

9. Conclusions

The impact of the sudden increase in online social network usage has yet to be fully understood. Society has been granted access to an unprecedented amount of shared data, most of which are publically available. From a research perspective, this information presents an opportunity to explore this relationship between online communication and real-world activities. The results of this exploration will set a foundation for using location-based data from online social networking sites in real-world socio-spatial research.

References

Ahas, R., & Mark, U., 2005, Location based services-new challenges for planning and public administration? *Futures*, *37*(6):547-561.

Backstrom, L., Sun, E., & Marlow, C., 2010, Find me if you can: improving geographical prediction with social and spatial proximity. *WWW 2010* pp.61-70

Barkuus, L., Dey, A., 2003, Location-Based Services for Mobile Telephony: a Study of Users' Privacy Concerns. *Proceedings of the INTERACT 2003, 9TH IFIP TC13 International Conference on Human-Computer Interaction. 9*

Bulgurcu, B., Cavusoglu, H., Benbasat, I., 2010, Understanding emergence and outcomes of information privacy concerns: A case of Facebook. *ICIS 2010 Proceedings.* 1:230.

Carrasco, J. A., & Miller, E. J., 2006, Exploring the propensity to perform social activities: a social network approach. *Transportation*, *33*(5): 463-480.

Chang, J., & Sun, E., 2011, Location 3: How Users Share and Respond to Location-Based Data on Social Networking Sites. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. pp.74-80

Cheng, Z., Caverlee, J., Lee, K., & Sui, D. Z., 2011, Exploring Millions of Footprints in Location Sharing Services. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. pp.81-88.

Crowley, D., 2012, Foursquare seeing 3 million check-ins daily. http://www.technolog.msnbc.msn.com/technology/technolog/foursquare-seeing-3-million-check-ins-daily-121720. Accessed May 2, 2012

Ellison, N.B., Steinfield, C., Lampe, C., 2007, The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication* 12(4) 1143-1168

Facebook Statistics, 2012, http://newsroom.fb.com/content/default.aspx?NewsAreaId=22. Accessed May 2, 2012

Kiefer, P., Raubal, M., & Schlieder, C., 2010, Time Geography Inverted : Recognizing Intentions in Space and Time. *18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS 2010), San Jose, California, USA,* pp. 510-513.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A., 2005, Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(33):11623-11628

Portes, A., 1998, Social Capital: Its Origins and Applications in Modern Sociology. *Annual Review of Sociology*. 24(1) 1-24