Activities in a New City: Itinerary Recommendation Based on User Similarity

Grant McKenzie, Krzysztof Janowicz

Department of Geography, University of California, Santa Barbara, CA, USA
{grant.mckenzie, jano}@geog.ucsb.edu

Abstract

Given the ubiquity of mobile devices, place recommender systems have begun to emerge in location-based service applications, taking advantage of progress in position technology. This evolution of location recommendation platforms has been augmented by the exponential growth of online social networks (OSN) leading to location-based social networks (LBSN) and local review sites such as *Foursquare* and *Yelp*. In this work-in-progress we propose a trip activity recommendation system based on the similarity between users of LBSN. Based on data gathered from multiple sources we demonstrate the early stages of a system that extracts the nuanced differences between *users* rather than just venues.

Background and Relevance

In recent years, the ways in which we think about data have shifted from standard generic search queries towards personalized recommendations. With advances in behavior tracking, recommender systems such as those employed by *Amazon* have emerged, organizing the plethora of content available online in a way that aims to more efficiently meet individual needs. Internet radio services such as *Pandora* and *Last.fm* monitor the listening habits of their users in order to filter new artists and create custom stations directly for their taste. Even television is beginning to move in that direction, companies like *Netflix* are using collaborative filtering methods to not only recommend movies, but also generate new content based on observed niche markets.

So far, research in the area of location-based recommender systems has primarily focused on GPS trajectories, previous check-ins from a single user or the favorite locations of social ties. In this work we propose to focus more on the similarities between individuals and the place-based decisions they make. By comparing multiple people to one another through the places they visit, one can expose latent location preferences inherent to certain individuals. Rather than simply recommending *Starbucks* to a coffee drinker, this model focuses on the nuanced properties of the coffee venues that the specific coffee enthusiast frequents. Given these preferences the model finds similar individuals that also value those properties and recommends places based on the related preferences of the similar individual. Simply exploring similarity on a Place of Interest (POI) by POI level masks the latent preferences that lead individual's to certain places. Stepping back and looking at the larger picture of the user as a whole, this innovative approach is able to recommend activity locations that speak to an individual's overarching values.

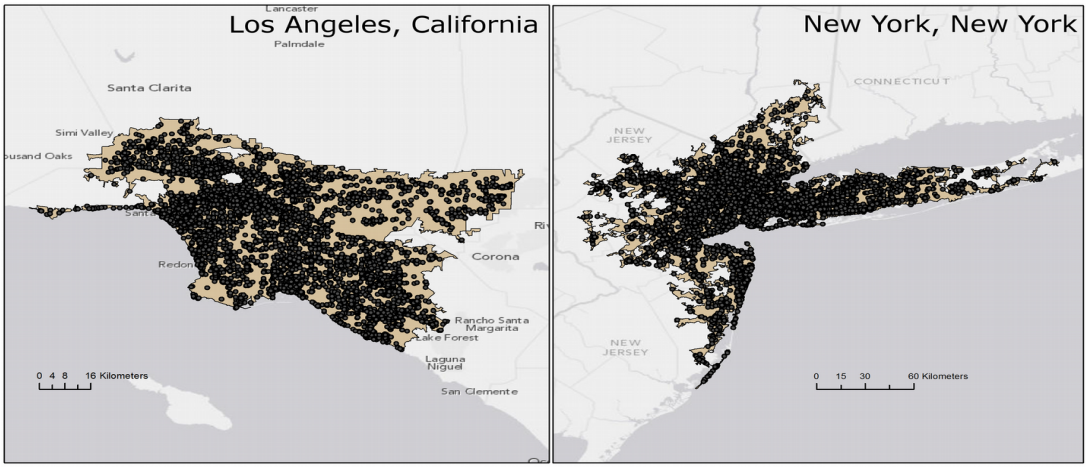
Previous work has focused on measuring user similarity through trajectory comparison. While some have just focused on the physical properties of a user's trajectory (Lee et al. 2007), others have taken a semantic approach to trajectory similarity measurement (Ye et al. 2011). Li et al. (2008) focused on hierarchical trajectory sequence matching to determine similar users. Their method made use of GPS tracks and stay points in which a user's activity was determined based

on the affordances of a specific location. Similarly, Ying et al. (Ying 2010) measured semantic similarity between user trajectories in order to developed a friend recommendation system. Recently, Lee and Chung (2011) presented a method for assessing user similarity based on LBSN data. While the authors also made use of check-in information, they used the hierarchy location categories supplied by *Foursquare* in conjunction with the frequency of check-ins to determine a measure of similarity. By comparison, our research is focused at the individual activity level, dramatically increasing the resolution at which user similarity is measured. Additionally, our approach is novel in that it makes use of an abundance of unstructured descriptive text (tips) in combination with semi-structured data provided by visitors of specific venues rather than a single categorical value.

Methods and Data

User Similarity

Building a model to measure similarity between users requires comparing users across a spectrum of properties related to the points of interest that they choose to visit. Built on our ongoing POI-matching work (McKenzie et al. 2013b), data queried from multiple local recommendation and review applications are combined to form the foundation from which a user similarity model is constructed. For prototyping purposes, 55,446 publicly available activity-based Foursquare check-ins were accessed for 538 unique users in the New York region and 247 in the LA region. Figure 1 shows the boundaries of the urban areas as well as the density and location of tweet/check-ins.



Provided venue identifiers via the shared check-ins, details for each of the venues was accessed via a single data Application Programming Interface (API). Descriptive information related to price, rating, number of "Likes," number of check-ins, unique number of users, category tags and unstructured review information were accessed for each of the 23,426 venues in our test set. While this information provides the foundation from which a user similarity model can be constructed, the data is still sparse. Using our POI-matching approach, 62% of venues from Foursquare were positively matched to venues in Yelp through the related API. These are good results given the known sparsity and partiality of online POI providers. The purpose of this matching is to add an additional layer of attributes and properties to enhance the existing content available through a single provider. This additional source lists over 30 different properties of a venue such as descriptive textual reviews and structured categories ranging from *Ambience* to *Wi-Fi* availability.

Trajectories

Organized temporally, user check-ins form trajectories that are often unique to a specific user. The difficulty with constructing trajectories lies in the sparsity of the check-in information users choose to share. Given a minimum of 30 check-ins per user, some form of aggregation is required to ensure feasible comparisons between users. In this work, check-ins were grouped in to either weekday (Monday - Friday) or weekend (Saturday & Sunday) activities based on check-in time-stamp. This division of check-ins follows the sensible notion that the types of activities one conducts during the week differ considerably from those conducted on the weekend. Aggregating activities in this way results in an average of 71.6 check-ins per weekday and 33.0 check-ins on the weekend.

Activity Clustering

Given the variety of activities conducted by different individuals, it is important to highlight the fact that not everyone follows the same daily activity pattern. Though the number of activities an individual conducts during the day ranges, recent research in time-use and activity behavior (Yoon et al. 2012; BLS 2012) lead us to group the check-in data in to nine daily activities. Based on this assumptions, *k*-means clustering analysis was applied to the check-in times with the purpose of clustering individual's daily activities in to nine user-specific clusters. Depending on the density of check-in times these clusters can vary radically between individuals, making the results of this research truly user-specific. Running *k*-means clustering for both the weekday and weekend activities separately produces eighteen distinct activity zones that can then be analyzed thematically before comparing themes across users.

To start, we choose a *Focal User* from a sample of trajectory sets and clustered her activities into eighteen distinct groupings. Each activity cluster is then buffered to produce a temporally adjacent set of clusters so that any randomly selected activity time can be assigned to one and only one cluster. This produces an *activity template* by which all other users' activity trajectories are clustered.

Common Properties within Clusters

In order to fully compare users, a rich set of POI properties is required. These properties define the POI, which in turn define the cluster and finally the individual user. The difficulty lies in the volatility of these properties within the POI dataset. For example, one POI may provide an *Ambience* value while another may not. Comparison of these two POI cannot rely on this attribute and the POI are reduced in their dimensionality, in turn reducing the strength of the comparison. Given the sparse nature of user-contributed data, this “missing property” issue is quite commonplace.

To mitigate this concern, the properties of each POI within an activity cluster are mined and a single representative value for each property across all venues is extracted. The use of common properties constructs what is often referred to as a *Prototype Activity*. The most common Ambience tag is stored as the Ambience attribute, the mean of the Foursquare rating and number of likes is stored and so on for all properties. Lastly, an aggregated topic signature is generated based on the average of each LDA constructed topic across venues that contained unstructured review text (Blei 2003; McKenzie et al. 2013a).

Prototypical POI

The *common* attribute values extracted from POI within a specific cluster are then used to fill in missing attributes of POI in the given cluster. For example, should one venue be missing a *Price* property, the most common price found for venues in the chosen cluster is assigned as the Price. While potentially not the *actually* Price for the venue, the *common price* reflects the common POI that the Focal User chooses to visit during a specific time of day. The inclusion of these common properties also increase the robustness of the user similarity model founded on POI attributes.

The next step in developing an activity signature for the Focal User is to extract prototypical POI that can be used to represent each cluster or "typical activity." To do this, each check-in within each cluster is compared to one another. A similarity value between each pair of POI is computed based on the attributes of the POI. In order to calculate the similarity, a modified *k*-nearest neighbor approach is taken by minimizing the dissimilarity values of each property in each POI. Not all properties of a venue are considered equal as some attributes do more to define a POI than others. For example the base category of a venue (e.g., Food, Entertainment, etc.) should be given more weight than the presence of Wi-Fi as the category is more indicative of the actual venue. In calculating the similarity of two POI, Equation 1 is applied to each pair of venues in any given cluster.

$$M_{sim} = (4JSD_Y + 4JSD_{FS} + 2Amb_Y + Price_Y + Price_{FS} + Rating_{FS} + Likes_{FS} + 4Cat_1 + 4Cat_2 + Cat_3 + WiFi_Y)/23 \quad (1)$$

As one can see, the amount of weight applied to each property ranges between 1 and 4. Unstructured content resulting in topics are weighted quite heavily as it has been shown that these topics alone do quite well at determining similarities between user trajectories (McKenzie et al. 2013a). Additionally, the first and second level Categories assigned to a POI are weighted heavily as the category of the location is quite important in assessing similarity. For example two POI may contain the same Ambience description of "casual" but one is categorized as a *Park* while the other is a *Bar*. Though both properties are important for defining the POI, the category of the venue should arguably be more influential. The resulting M_{sim} variable represents a dissimilarity value for each pair of POI in each cluster. This comparison value is averaged for each individual POI and ranked from lowest to highest. The POI that shows the minimum value (highest similarity to all other venues) is labeled the *prototype POI* and is used to define the typical activity location attended by the Focal user in a specified time frame.

Comparing Users

Once the prototypical POI for the Focal User's clusters are determined (e.g., Los Angeles user), the same must be done for all users in the destination (e.g., New York) sample set. The *activity template* defined by the clusters of the Focal User is applied to each user trajectory in the destination sample set. This temporally clusters the activity of each user by the time frames created for the Focal User. Following the previously stated procedures, the common POI properties within each cluster are obtained and applied to the vacancies in the individual POI. These venues are then compared within each cluster and prototypical POI are extracted and assigned to each cluster. Using this method each user in the dataset is characterized as a temporal series of eighteen activities evenly split over two days (week day and week end).

Provided the Focal User and the set of destination users, each of the Focal User's eighteen prototypical POI are compared with the related POI in each of the destination users' clusters. The comparison method follows Equation 1. The Jensen-Shannon divergence metric is used to

calculate dissimilarity between Foursquare topics in each venue pair and the same is done for Yelp topics. Nominal values such as Ambience and category levels are calculated based on the number of term matches that are found between POI (e.g., “Casual, Fine-dining” and “Hipster, Casual” equals 0.5). Ordinal and ratio values such as price, rating and number of likes are calculated as the absolute value of the difference between POI properties normalized by the maximum property value. In the case of likes, this number is already normalized by the total number of check-ins at the given Foursquare venue. Finally Wi-Fi is a simple boolean value that either matches or does not.

In this way, a dissimilarity value is calculated for the combination of each of the prototype activities in the Focal User's trajectory and each of the destination users' prototype activities. These dissimilarity values are averaged by user to produce a single value for each user. The destination user that is said to be most similar to the Focal user is the one with the smallest average dissimilarity number.

Conclusions and Next Steps

While the above methods describe a novel system for assessing and ranking activity similarity between users, the methods must be applied and tested. One possible application lies in a tip-planning system for recommending locations and activities to out-of-town visitors. Provided an individual's trajectory, an itinerary would be generated for that person in a foreign city, based purely on the activities of similar individuals in the chosen city. In order to test this application, a study will be conducted in which a number of participants are presented with a recommended itinerary (based on the most similar user's activities) along with itineraries based on randomly selected users. Participants will be asked to indicate which of the itineraries they would most likely choose given their daily activity pattern.

References

Blei, DM., Ng, AY., Jordan, MI. (2003) Latent dirichlet allocation. *Journal of Machine Learning Research* Volume 3. pp. 993–1022

BLS - Bureau of Labor Statistics (2012) American Time Use Survey Summary (USDL-12-1246) URL: *http://www.bls.gov/news.release/atus.nro.htm*

Lee, J., Han, J., Whang, K. (2007) Trajectory Clustering : A Partition-and-Group Framework . In *International Conference on Management of Data*, pp 593–604.

Lee, M., Chung C. (2011) A user similarity calculation based on the location for social network services. *DASFAA*, pp. 38–52.

Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W. (2008) Mining user similarity based on location history. *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems - GIS ’08*, pp 1.

McKenzie, G, Adams, B., Janowicz, K. (2013a) A thematic approach to user similarity built on geosocial check-ins. In *Proceedings of the 2013 AGILE Conference,* Springer

McKenzie, G., Janowicz, K., Adams, B. (2013b) Weighted multi-attribute matching of user-generated points of interest. *In ACM SIGSPATIAL 2013* Short Paper

Ye, M., Shou, D., Lee, W., Yin, P., Janowicz, K. (2011) On the semantic annotation of places in location-based social networks. *In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge discovery and data mining*, pp. 520–528.

Ying, J., Lu, E., Lee, W., Weng, T., Tseng, V. (2010) Mining user similarity from semantic trajectories. *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks - LBSN '10*, pp. 19

Yoon, S., Ravulaparthy, S., Goulia, K. (2012) Dynamic diurnal social taxonomy of urban environments. *13Th International Conference on Travel Behavior Research.*