Personalized Location Recommendations in LBSNs (Location-based-Social-Networks)

Shubham Mittal (104774903)

February 2017

1 Problem & Goal

With the recent technological advancements, social network websites today, are able to deeply connect with the users by allowing users to share locations by "checking-in". This has bridged the gap between the physical and digital world and enabled a deeper understanding of user behavior and preferences. A location based social network is different from a traditional social network in the sense that it not only adds a geographical layer to the existing social layer, but also a content layer consisting of tips, photos etc. creating a rich dense graph, with a huge potential for mining interesting information. Although there have been many advancements in the recommendation system for online social networks, recommendations for LBSNs haven't been investigated to a great detail in the academic literature.

My goal in this project is to develop a personalized location recommendation system for users based on the user "check-in" data and tips/reviews about the place. These data are heterogeneous in nature, but both describe the user's preferences for different venues. I plan to tap into this heterogeneous data stream, perform expertise search, compute the ranking and rating of the place and finally match it with the user's own review to obtain a personalized, ranked list of locations for the user. This has tremendous applications and provides a useful metric for both the club/restaurant owners as well the users.

I also plan to integrate the user's social relationship and his location behaviour to give accurate location recommendations. Such heterogeneous tapping of data is very limited in academic literature especially in the scope of LBSNs. I expect to see accurate ranked and personalized location recommendations for users based on their social circle, past "check-ins" and generic reviews of the venue. At the end, each user will see the top-N new venues matching his interests.

2 Data Plan

For both training and testing, I plan to use the publicly available Gowalla dataset. Gowalla was a popular location-based social network having more than 600,000 users and was shut down after being acquired by Facebook in late 2011. The dataset contains 36,001,959 "check-ins" made by 319,063 users over 2,844,076 locations. Locations are tagged into seven different categories like Entertainment, Food, Nightlife, Outdoors, Shopping and Travel. Each of these categories consist of many subcategories. However, the Gowalla dataset does not contain tips/feedback information from the users.

Another popular dataset is the NYC Restaurant Rich Dataset which includes check-in, tip and tag information of restaurant venues in NYC collected from Foursquare, another popular location-based service. The dataset contains 3112 users and 3298 venues with 27149 check-ins and 10377 tips. However, this particular dataset does not have the user social relationship data.

I plan to work on my initial models using these two datasets and continue to search for a complete rich dataset. If such a dataset cannot be found, I'll present the results of the models separately for these two datasets.

3 Solution Plan

My first challenge given the dataset would be to find the popular locations by taking inputs from the entire set of users. Users may have variable feedback and check-in count at different venues. Some obvious features

in deciding the popularity would be the check-in count, count of unique visitors etc. I'll have to study each of the different metrics and find out if it affects the popularity of venue or not.
The second challenge would be to mine relevant text in the feedback/tips from restaurants and decide whether the user favors the venue or not. In the current form, the dataset contains reviews in the form of broken English slangs, for example: *"total madhouse and that's because the french onion soup keeps everyone coming back.. oh, and celebrity sightings too. leave room for dessert, my 2 faves? Tarte Aux Pommes Crêpes Suzette."* Also, the review could by totally unrelated like *"In 2011, 60% of the 25,000 orange, red and white light bulbs will be LED. The conversion to LED will be complete by 2013 and will save Texas Tech an estimated 80% on electricity costs."* Another challenge with this dataset is that it may contain empty tags for a venue since none of the users ever labelled it. To counter this, I will have to first preprocess the data to fill out any missing tags (or ignore that data entity) and next extract a score or a binary preference value from the user's feedback.

I would first tackle the venue popularity problem and try to deduce the most popular venue based on user's opinions and other features. I plan to complete this in two phases - a) finding the popularity of a venue, and b) finding the popularity of the category (tag). I will do this by first computing the frequency by which a user comes back to the same venue and extract the relevant tag from that venue to identify the user's maximal interests. I would then compute the category popularity by grouping the data based on the categories. My next step in the solution would be to identify the user social circle and then study its effects on his venue preferences. I will use the existing algorithm to adopt a graphical version of LDA into the check-in dataset for the geographic topic modeling. My final work would be implementing the recommendation system, so as to recommend personalized venues to users. I'll use the existing POI recommendation problem for this part of the solution.

4 Evaluation Plan:

I will use the standard precision and recall metrics to evaluate the approach.

5 Project Time Line:

I feel this project has the potential to easily span over multiple months, and to restrict to a course level project, I'll apply the basic data mining functions at each step. The project time line would look something like this:

- Relevant Research, Approach and Dataset Finalization - Feb 20th
- Venue popularity implementation - Feb 27th
- Recommendation engine implementation - Mar 9th
- Evaluation of the implemented algorithm - Mar 14th
- Presentation and project completion - March 17th