User profiling

In case of our recommender system there might be two different types of users for which we are providing recommendations:

1) New user, inputting it's preference through the website. There are several ways of identifying user's preferences:
    a. Requesting user to fill out specific features he is interested in. This approach is feasible and we will implement it first. One drawback is that it can be a burden for a user to fill out all the necessary features. Thus, in this case, the features of the venue must be selected such that they are precise enough for good recommendations and still general enough for easy specification.
    b. Another approach that could be taken is to ask user to rate several venues of different characteristics and by using the ratings, create the user profile

2) Existing user. This is where user data from Foursquare comes into play. While I think that this approach is interesting, it is much more important to implement the recommender for the first type of users. Here is the analysis of the data available that could be relevant:

Field	Description	Complete Object	Comments
id	A unique identifier for this user.	✓	Important unique id.
friends	Contains count of friends for this user and groups of users who are friends. Right now will only contain a group where type is friends, but it's subject to change. Groups are omitted when viewing self The full set of friends is at users/XXX/friends.	○	By accessing the friends list we can create the social graph of connections. This could be used for recommending venues to existing users based on their friends or even to new users, based on the people that had interactions with that venue and their friends.
type	Present for non-standard user types. One of page, chain, celebrity, or venuePage. Pages are brand pages, such as Bravo, chains are pages that own a set of venues, like Starbucks, celebrities are users that other users can follow, like Mario Batali, and venuePages represent single venues creating content, such as My ArenA creating updates. venuePage's do not really have a	○	This is not important at the moment, but I left it here, because celebrity type might be interesting. We should though check on how many of fetched users have the type attribute.

Field	Description	Complete Object	Comments
	user profile, show the veue page instead.		
homeCity	User's home city	✓	Can be interesting, but by observing the dataset it is notable that people do not enter this data in a structured way. For example one might write “NY”, while someone else writes “New York” etc.
gender	male, female, or none	✓	This is important in determining the male/female ratio in the venue.
contact	An object containing none, some, or all of twitter, facebook, email, and phone. Both are strings.	✓	Some additional data could be inferred about the user from here, but from what I have seen in the dataset, this contact information is very rare.
bio	A short bio for the user.	○	Usually empty, would require NLP.
tips	In compact users, if present, contains the count of tips from this user. In full users, contains count and an array of selected tips as items, which may or may not be empty. Full set of items at users/XXX/tips	○	Tips are one of the users interactions with venues. We can model check-ins by them. Additionally, some NLP could be applied to the tips text, but I don’t think that it’s feasible given the timeframe.
lists	If this user has lists, *groups* contains *created* for lists created by this user. In compact, just the count is provided, and in full, some sample lists are included. In full form, lists also contains a count of the lists created and followed by this user. Full set of items at users/XXX/lists	○	Lists could be interesting because they are a sort of user generated recommendations. They are somewhat rare in the dataset at first sight, but could certainly be used for enriching the user – venue connections. We could for example add all of the venues on the lists for the given user to its profile and rank them by appearance in these lists (lists are usually subjectively sorted by their creators and contributors), take a look at https://foursquare.com/timeoutnewyork/list/best-new-burgers-in-nyc
followers	If this user can have followers (i.e. it's a celebrity, venue, or page), contains count of	○	This is can be used for additional linking between users and venues, giving more weight to their interactions, but this is

Field	Description	Complete Object	Comments
	followers for this user in compact, and, in detail, groups of users who follow this user, split into friends and others.		accessible through venue API and not important in terms of user profiling.
following	groups of pages this user has followed.	✓	This is another way to get additional lists and tips. Take a look at https://foursquare.com/corcoran_group for an example. This can get very deep so we need to decide together what data do we want to exploit.
badges	Contains the count of badges for this user. May eventually contain some selected badges. Use users/XX/badges to get the list of badges.	✓	This doesn't mean much to us as there is no relationship between badges and venues. It could though be used as a factor in evaluating user relevance.
mayorships	Contains the count of mayorships for this user and an items array that for now is empty. Use users/XXX/mayorships to get actual mayorships.	✓	This is an important piece of information. It is an additional source of data for user-venue interactions, but must be weighted more strongly than tips and photos. It additionally can be a factor in determining user's relevance.
photos	Contains the count of photos this user has. May contain an array of selected photos as items, full set of items at users/XXX/photos, but only visible to self.	✓	This is very valuable information. Among other data, will model interactions with venues based on photos and could infer some user relations based on likes on photos.
scores	Contains recent, max, checkinsCount, and goal for showing user's current game status.	✓	Could be used as an additional factor in determining user's relevance.
checkins	Contains the count of checkins by this user. May contain the most recent checkin as an array items containing a single element, if the user is a friend. Users can fetch users/XXX/checkins for their own complete history.	✓	Since we do not have access to the check-ins themselves, this could be

There is a lot of data available on users, although sadly we do not have access to check-ins.

From what I have concluded, this data can be used in majority:

- To model user – venue interactions and by aggregating the venues, extract the features that will be compared for similarity against our venue dataset.
- To asses user relevance. For example, when we are trying to make recommendations to an existing user, venues visited by the most relevant of his friends will have higher chance of being recommended. Or if a very relevant user had interaction with some venue, that venue might get a higher score to some feature.
- Additionally, the data could be used to create a social graph and to do collaborative filtering, but as agreed we leave this out of the scope of this project for now.