Jay Hack, Ankit Kumar and Sam Beder
jhack@stanford.edu, ankitk@stanford.edu, sbeder@stanford.edu

Calendar Event Recommendation System Proposal

I. Motivation

SpotOn is valuable for users because it helps them manage their pre-planned activities. We believe SpotOn can leverage the enormous amount of data they have gathered about their user base to recommend events that may be of interest to them; informed recommendation benefits both the user, who see events that actually like, as well as the event planners, who gain more traffic at their events. Therefore, we propose to build a system that (1) models SpotOn users and their interests and (2) applies this knowledge to recommend events.

II. Intended Approach

As a team, we have solved several data science problems that are similar to user event recommendation. Below are two ideas that we have brainstormed for our initial attempts to create a robust event recommendation system. Please note, however, that this is by no means an exhaustive list of what we intend to try; as data science is an empirical pursuit, we recognize that our initial approach may need to be modified or exchanged for other techniques to build a successful recommendation system.

Leveraging Natural Language Descriptions via Topic Modelling:

The natural language descriptions of calendar events that SpotOn has access to are one of the richest and most informative sources of information available. We propose to utilize this information in order to come to a subjective account of the content of calendar events and, subsequently, users' interests. In particular, we intend to use latent dirichlet allocation (LDA) in order to (1) characterize all events in terms of the 'topics' that they are composed of, then (2) characterize a user in terms of the topics of the events that they have attended in the past. As a team, we have applied a similar procedure with great success to the problem of matching users for anonymous online chats (using data from Chatous.com, see [2]) and expect it to yield similar results in this domain due to the problem similarity.

Latent dirichlet allocation is an unsupervised learning algorithm by which each element in a set of 'documents' (i.e. calendar events) is described as a compilation of a set of 'topics,' each with a given weight. These 'topics,' which are also learned through an unsupervised learning algorithm, tend to be surprisingly good approximations of real-world topics of discussion. For example, see [1] for a discussion of the LDA topics derived from the set of all wikipedia articles. One can see that words in a single 'topic' have many semantic similarities, such as *river, lake, island* and *mountain*. Once each individual event can be characterized in terms of its constituent LDA topics, an individual

user can be represented as a weighted sum of each of the LDA topics of the events that they have attended in the past. We have found that this user representation, a vector with length *n*, where *n* is the number of topics, is useful in finding subjective similarities between two users as well as in recommending entities for the user to experience in the future. (In particular, we found that a user's LDA topics were highly informative features for user compatibility matching using Chatous.com data.) For a more technical, in-depth explanation of LDA applied to user characterization and recommendation tasks, please see our previous paper, [2].

Collaborative Filtering and Network Analysis:

In addition to natural language data, we expect that the network structure induced by SpotOn users and events that they have attended will be highly informative. SpotOn's calendar data induces both (1) a bipartite network, in which users connect to events they have attended, and (2) an undirected network, where events are connected by edges representing mutual attendees.

The bipartite network allows us to employ collaborative filtering (CF) techniques, which attempt to characterize both users and events in terms of their mutual similarities, then predict new user-event pairings using a similar approach. In particular, we intend to explore CF techniques that utilize k-nearest neighbors due to their past success in similar tasks. In addition, previous attempts at solving similar problems have successfully used temporal notions of both user bias and event popularity. For a more in-depth treatment of the CF techniques mentioned here, see the winning Netflix Prize paper [3]. We note that the Netflix data set and the SpotOn data set, while somewhat similar, also have significant differences, and as such the techniques employed in the paper will need to be adapted or changed completely.

The undirected network between events will allow us to use community detection techniques to output event clusters, and will subsequently allow us to characterize users in terms of which event clusters they frequent. Given the size of the network, heuristics that prioritize computation time such as the Louvain Method [4] will likely be the algorithm of choice; however, clustering sub-networks with more precise methods is a technique to explore. Our team has extensive experience with community detection algorithms, in particular [5], and some of the ideas in the algorithm we developed in [5] would likely be applicable here.

While both approaches described above (LDA- and network-based) have a similar goal, namely finding similarities between events, note that they provide fundamentally different accounts of event similarity. While LDA attempts to find a subjective characterization of an event's content through the topics mentioned in its description, network clustering approaches find event similarity based on a the more objective commonality of mutual attendees. Therefore, we expect that they will offer complementary insights into event similarity and are worth implementing side-by-side.

III. First Iteration Deliverable

As the first iteration in our design process, we will deliver a module that serves of a 'proof of concept' for how event recommendation can be integrated into SpotOn's existing platform. Specifically, we will deliver a discrete python module that produces a sorted list of recommended activities for each user, along with a confidence score for each recommendation, given a file containing JSON representations of users' calendar events and recommended activities (the format initially provided via email). We will provide full documentation of how to use the module as well as a description of the algorithms used internally. Furthermore, we would like to give an in-person demonstration of our module and the underlying algorithms' capabilities. After completing this module, we will move on towards actually integrating it into the current SpotOn system.

IV. Compensation

We propose a two-tiered framework for compensation. First, our compensation should track the success of our program, such that we are paid in proportion to the amount of value added to SpotOn, as measured by some agreed-upon evaluation metric. (Users' click-through rate or total number of clicks on recommended links, for example). Second, because it is relatively difficult to evaluate our algorithm during development, we would like to be partially compensated with a flat fee. We guarantee a working version of the first iteration deliverable, described above, as well as our services as data science consultants--these may include, but aren't necessarily limited to, thorough descriptions of the approaches we tried, what worked, as well as general approaches for developing calendar event recommendation with user data. We believe our work will be valuable for SpotOn, both for the core functionality of the module we develop, and for contributing to data science competency of firm, generally.

In closing, we'd like to recognize that the terms above represent a first offer, and would propose meeting in person to discuss the project and compensation in greater detail. We are generally flexible with respect to scheduling and are very excited to start working on this project.

References:

[1] Rehurek, Radim. "Experiments on the English Wikipedia." January 2014.
http://radimrehurek.com/gensim/wiki.html

[2] Hack, Beder, and Zamoshchin. "Predicting User Compatibility Online with Advanced Natural Language and Network Features." December 2013.
http://www.stanford.edu/class/cs224w/projects2013/cs224w-009-final.pdf

[3] Koren, Yehuda et al. "The BellKor Solution to the Netflix Grand Prize." August 2009.
http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf

[4] Blonden, Vincent et al. “Fast unfolding of communities in large networks.” March 2008. http://arxiv.org/pdf/0803.0476v2.pdf
[5] Kumar, Ankit. “Detecting communities via Absorbing Markov Chains.” December 2013. http://www.stanford.edu/class/cs224w/projects2013/cs224w-065-final.pdf