CS54701: Final Project Proposal

Brian Olsen

1 Proposal

1.1 Problem Statement

Spatial information retrieval (SIR) is an extension of the well studied information retrieval topic which focuses on evaluating a user query and returning documents that satisfy this user's information need. SIR aims to accomplish the same objective but with regard to spatial proximity specified in the query of the user. An increasingly popular technique called collaborative filtering has gained much popularity with the increasing e-commerce applications that drive a personalized experience for the user. Collaborative filtering uses prior data from similar users to suggest possible items that the user has not tried yet. In the same manner we have similar types of recommendation systems for spatial systems today that track multiple users spatiotemporal activity through "check-in" data via micro-blog sites such as Twitter and Foursquare.

Example: Consider the following scenario where a user is going out on a Saturday night and opens up the app to look for something to do that night. If the user tends to go to clubs on Saturday evenings then that user will probably have other users that do the same thing at the same time. However maybe the other users are in a different part of town and now the user get's notified of new clubs across town that would have otherwise been unknown to the user.

1.2 How is it done today:

In a recent paper published in the IEEE International Conference on Data Mining, the STT (spatio-temporal topic) model was suggested as an approach to spatio-temporal collaborative filtering. Their implementation uses the Latent Dirichlet Allocation (LDA) to assign different locations into certain topics or themes. Some areas may be business oriented while others entertainment. Once the themes are established an Expectation Maximization probability is applied to the users and items an each gets a probability that will then determine a ranking that will be used to display to the user.

1.3 My Approach:

My approach is that instead of using a statistical model I plan on applying a collocation pattern to the data. Recently there have been many advances in spatial data mining that seek to find collocation patterns using approaches that piggyback off of association-pattern mining. While this has been applied to non-spatial collaborative filtering it has yet to be applied in spatial collaborative filtering. I feel confident that this approach will work as the research for the association rules applied to spatial data have already been realized. For instance, we know that spatial data is not anti-monotonous and this ruins fundamental pruning advantages from traditional pattern mining. We can apply the current research using a modified pattern mining approach. I will finally rank the items by the one's with

either highest support, confidence or both. It is hard to say if this approach will be more effective but I believe since pruning is a key element in this algorithm we may be at the least more proficient in the computation than the mentioned paper above.

1.4 Challenges:

- Assigning regions or locations to topics. (multiple topics? how is this done?) My proposal to defeat this challenge is to split the locations up into categories based on openstreetmap data or possibly to use LDA similar to the paper.
- Users tend to travel to different cities and if they are generally in one city A and travel to city B their suggestion may have an affinity towards recommending locations in city A even when the user is in city B.
- Getting data tagged using openstreetmaps may prove to be difficult and inaccurate due to collaborative data source.

1.5 Timeline:

1. April 1 - Obtain all data sources (Twitter, openstreetmap, etc...)
2. April 8 - Preprocess and tag all data with categories (using LDA or openstreetmaps)
3. April 15 - Finish algorithm implementation of both models (Mine and theirs).
4. April 22 - Finish testing both models.
5. April 29 - Finish writeups and submit project.

1.6 Evaluation:

What I believe defines success is if I can successfully set up the data, algorithms and good objective tests to verify whether or not my idea has any better qualities over the current implementation. As described in class previously the objective isn't to come up with an idea that must be better but that I simply obtain objective data one way or another and clearly state my methods of running my experiment.

Algorithm Implementation and Data	Write-up(3-5 pages)
40%	60%