Spatio-Temporal Site Recommendation

Blinded for Double-Blind review

***Abstract*—Recommendation systems have become extremely common in recent years, and are utilized in a variety of areas to predict the "rating" or "preference" that a user would give to a point of interest (PoI), such as a restaurant, a hotel, or a bar. Such systems typically produce a list of recommendations by considering previous ratings of the user, as well as ratings of other users. Not every person rates every point of interest they visit. In this work, we want to explore the use of spatio-temporal data to improve recommendation systems: We postulate that spatio-temporal user data may indicate the liking or disliking of a point of interest. Clearly, if a user frequently visits the same PoI, stays at the PoI for long times, and is willing to travel a long distance to visit a PoI, that might indicate that user likes that PoI. Thus, we propose to extract user-PoI relation features from spatio-temporal trajectory data only. Using these features, we use out-of-the-box data mining and machine learning solutions, to estimate the popularity of a PoI. Our experimental evaluation shows, that the features extracted from spatio-temporal data able to accurately predict the popularity of a PoI, using ground-truth data from Yelp as a baseline.**

I. Introduction

Modern technology to capture geo-spatial information produces a huge flood of individual trajectory data, coupled with a new user mentality of utilizing this technology to voluntarily share information. By mining this data, and thus turning it into actionable information, the McKinsey Global Institute [3] projects a "$600 billion potential annual consumer surplus from using personal location data globally". Towards this goal of making location data actionable, we propose to mine potential user ratings for location sites (e.g., restaurants, shops) from spatial-temporal data. User rating is critical in many applications, especially recommendation system. Techniques such as widely used collaborative filtering use known user rating information to estimate the interests of a user. However, in practice, the user-site rating matrix is usually highly sparse, meaning that there are not enough user rating information to perform such tasks. This is know as the cold-start problem. There are many reasons, for example, many users do not want to spend time to rate locations they visited, or locations such as shops do not have an efficient way to ask users for feedback.

To solve this problem, we propose to obtain implicit user-rating rather than explicit user ratings. Our approach uses trajectory data to find users that have likely visited a site. Since this data does not explicitly tell us whether the user likes that site, we propose to mine features from the users trajectory, which we intuitively expect to implicitly describe whether the user likes that site. The features that we propose to obtain from trajectory data include:

- **The frequency of visits of a user.** If a user visits a site only once in their life, chances are that the user did not like the site enough to return. If the site if frequented often by the user, he seems to like it.
- **The length of stay of a user.** If a user stays at a restaurant or a hotel for a long time, that indicates that he likes the site.
- **The distance to their home base.** If a user is willing to take a long journey to reach a site, thus bypassing other, similar sites, that indicates that the user is subject to a strong attraction from that site, indicating that the user likes that site.

There are several advantages of using location data for location rating: (i) A potential solution to the cold-start problem, (ii) no user effort to capture their recommendation such as filling in rating forms, and (iii) it is based on user's behavior, thus more objective and prone to alteration by fake-user-ratings and spam/bot-user-ratings.

Our approach becomes viable, to due to the abundance of large open-sources collections of voluntarily contributed trajectory data, including the following data sources:

- **Location-Based Social Networks (LBSNs)** allows user to "Check-in" into a physical site such as a hotel, a restaurant or a metro station. Fairly large LBSN datasets, made anonymous, are available publicly. For instance, the FourSquare dataset used in [6] contains more than 30 million checkins and is available publicly. Such data explicitly includes the sites that a user has visited.
- **Geocoded Social Media Data:** is obtainable from public streaming APIs including for Twitter, Instagram, and Flickr. These data sources provide low-frequency trajectory data. Yet, microblogs and images are often published in sites of interest to a user, thus giving implicit information about the users site preferences.
- As part of the effort to create **Open-Street-Map (OSM)** [4], [1] road network data, users have been uploading GPS traces of their routes to the OSM site. While these routes are typically used to digitize the road network, the majority of routes is uploaded by pedestrians, and can be used as an indicator of sites that the corresponding user frequents. These trajectories are publicly available through the OSM API.

To describe our approach of site-recommendation using trajectory data, the rest of this work is organized as follows. We survey the state-of-the-art on site-recommendation and location-based recommendation systems in Section II. Then, we formally define the problem location-based site recommendation in Section III. Our solution, using deep-learning to bridge the gap from user-behaviour to user-site-recommendations is given in Section V. Our solution is evaluated in Section VII, showing that our rating prediction for a restaurant is able to closely predict authoritative ground-truth site-ratings obtained from Yelp. We conclude our work in Section VIII.

II. Related Work

III. Problem Definition

In this section we formally define a user-trajectory, and define our notion of a user-site-stay, which we are going to use to extract features to estimate the affinity between a user and a site later in Section V. We first start by defining a trajectory as follows.

Definition 1 (Spatio-Temporal Database): Let $\mathcal{U}$ denote a set of unique user identifiers, let $\mathcal{G} = [-90, 90] \times [-180, 180]$ denote the space of longitude/latitude geo-coordinates, and the $\mathcal{T}$ denote the time domain. A *spatio-temporal database* $\mathcal{ST} \subseteq \mathcal{U} \times \mathcal{G} \times \mathcal{T}$ is a collection of triples $(id \in \mathcal{U}, (lat, long) \in \mathcal{G}, t \in \mathcal{T})$. Each triple $(u, s, t) \in \mathcal{ST}$ is called an observation.

We group a spatio-temporal database into observations of the same user, denoted as user-trajectory, formally:

Definition 2 (User-Trajectory): Let $\mathcal{ST}$ be a spatio-temporal database and let $u \in \mathcal{U}$ be a user. The set

$$\mathcal{ST}(u) := \{(u', (lat, long), t) \in \mathcal{ST} | u = u'\}$$

is called the user-trajectory of user u.

In order to obtain recommendation information from a user-trajectory, we need to link the user-trajectory to sites, such as restaurants and hotels. For this purpose, we join a spatio-temporal database with a database of points of interest (such as provided by Open-Street Map) like restaurants and hotels. Next, we define our notion of a *stay*. A stay is an event of user visiting a site, enriched by the duration of the stay.

Definition 3 (Stay Trajectory): Let $\mathcal{ST}$ be a spatio-temporal database and let $\mathcal{S} \subseteq \mathcal{G}$ be a collection of $(lat, long)$ pairs of sites. A stay is a triple $(u \in U, s \in \mathcal{S}, (t_{\text{start}}, t_{\text{end}}) \in T \times T)$, indicating that user u has stayed at site s from time t_{start} to time t_{end}. We let $(\mathcal{ST} \bowtie \mathcal{S})$ denote the set of all stays mined from all trajectories $\mathcal{ST}$ using all sites in $\mathcal{S}$. The sequence of all stays a user $u \in \mathcal{U}$ is called the stay-trajectory $(\mathcal{ST} \bowtie \mathcal{S})(u)$ of u, defined as:

$$(\mathcal{ST} \bowtie \mathcal{S})(u) := \{(u, p, (t_{\text{start}}, t_{\text{end}})) \in \mathcal{S} | u = u'\}$$

Finding stay points in trajectory and PoI databases is a research topic that has raised attention in the past. For instance, a state-of-the-art approach [2], [8], [7], [5] uses a distance threshold θ_d and defines a stay as the duration of time where the trajectory does not exceed a distance of θ_d to a PoI. In this work, we assume that *stay* detection algorithms are already applied to a trajectory, such that user-trajectory is mapped to a sequence of PoI stays. This assumption is discussed in Section **??**.

Given stay-trajectories for each user, the challenge of this work is to predict the rating between users and a site. As site is PoI which can be rated by a user, such as a restaurant or a hotel. We assume that we have a recommendation database, where users can rate sites. We assume normalized rating values in the interval $[0, 1]$, where 0 corresponds to the lowest rating and 1 corresponds to the highest rating.

Definition 4 (User-Site Recommendation Database): A recommendation database $\mathcal{R}$ is a set of user ratings $\mathcal{R} \subseteq \mathcal{U} \times \mathcal{S} \times [0, 1]$.

The task of this work is to predict the triples in R. That is, the challenge is to predict the rating that a user $u \in \mathcal{U}$ will give to a site $s \in \mathcal{S}$, using a spatio-temporal given in $\mathcal{ST}$.

IV. Discussion: Stay Point Detection

V. Spatio-Temporal User-Site Feature Extraction

VI. User-Site Rating Prediction

VII. Experimental Evaluation

VIII. Conclusions

References

[1] M. Haklay and P. Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, October-December:12–18, 2008.

[2] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 34. ACM, 2008.

[3] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. 2011.

[4] Open street map. http://www.openstreetmap.org.

[5] X. Xiao, Y. Zheng, Q. Luo, and X. Xie. Finding similar users using category-based location history. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 442–445. ACM, 2010.

[6] D. Yang, D. Zhang, L. Chen, and B. Qu. Nationtelescope: Monitoring and visualizing large-scale collective behavior in lbsns. *Journal of Network and Computer Applications*, 55:170–180, 2015.

[7] Y. Zheng, X. Xie, and W.-Y. Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

[8] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800. ACM, 2009.