Want a Coffee? Predicting Users' Trails

Wen Li
Delft University of Technology
Delft, Netherlands
wen.li@tudelft.nl

Carsten Eickhoff
Delft University of Technology
Delft, Netherlands
c.eickhoff@tudelft.nl

Arjen P. de Vries
CWI
Amsterdam, Netherlands
arjen@acm.org

ABSTRACT

Twitter and Foursquare are two well-connected platforms for sharing information where growing numbers of users post location-related messages. In contrast to the longitude-latitude geotags commonly used online, e.g., on photos and tweets, new place-tags containing category information show more human-readable high-level information rather than a pair of coordinates. This grants an opportunity for better understanding users' physical locations which can be used as context to facilitate other applications, e.g., location context-aware advertisement. In this paper, we verify the assumption that users' current trails contain cues of their future routes. The results from the preliminary experiments show promising performance of a basic Markov Chain-based model.

Categories: H.2.8 [Database Management] Database applications - Data mining J.4 [Computer Application] Social and Behavioral Science

General Terms: Experimentation

Keywords: Twitter, location-based estimation, text mining, geotagging, mobile applications

1. INTRODUCTION

Twitter is a platform for sharing various kinds of information, such as photos, videos and even users' geo-locations. This is going far beyond what the service was designed for; sending text messages of 140 characters. Originally, user messages, so-called tweets, only contained a pair of geo-coordinates to indicate the location where the user issued his/her tweet. This hardly provides any information about the environment around users at the time of posting. Last year, Twitter extended its API to supporting *Places of Interest* (POI) to facilitate users tagging their tweets with more accurate high-level geo information. This novel kind of geo-tags provides several properties such as names, bounding-boxes, and information about geographically-related entities, giving easy access to the functional aspect of locations.

Foursquare is a dedicated geo-location sharing platform. Users are encouraged to check-in at a POI, that is, to post a message on both Twitter and Foursquare, declaring their visit, upon which they are rewarded with badges, coupons, etc. The service's API provides rich information, such as POI categories and URLs to homepages. This kind of information enables making connections between users, locations and their potential activities. The place level geo-information arising in Twitter and Foursquare facilitates our research on predicting users' likelihoods of visiting a place based on their current and previous visits. With this kind of knowledge, advertisers could display targeted information more accurately to earn more attention and clicks. For example, having known a user would go for a coffee after work, a suggestion of an attractive lately-opened café would be more persuasive before he/she already visited one.

Recently, there were many discoveries in the domain of location-related social information. Backstrom et al. [1] studied the vocabulary gap between users in different areas in the US and proposed a method based on local word filtering to estimate users' hometowns. Kurashima et al. [2] proposed a method of integrating Markov Chains and Topic Models to recommend travel routes based on geo-tagged photos shared online. Sadilek et al. [3] studied the relation between friendships and location visiting patterns and proposed a friend recommendation method. These studies on location-aware applications show an interesting connection between a user's location and his/her characteristics in other aspects.

In this paper, we formalize our task as, given a user's history of visits in a trail, how can we estimate the types of POI in the future part of trail? Here, a trail is defined as a sequence of POIs a user visits during one day. We experimentally validate whether we can predict user behavior represented by the POI type of their next visit based on data from Twitter and Foursquare using a Markov Chain-based model. The method can also be easily integrated into applications on mobile devices which may facilitate a location recommendation or context-aware advertising.

2. DATA

The data were collected via Twitter's API following a strategy of tracking active users of POI-tagged tweets and the places popular among those users. After 6 months of consecutive crawling, we hold a corpus of 1 277 596 tweets tagged with 236 369 unique POIs. In Figure 1, we can see that the distribution of tweets over POIs approximate a power law distribution and around 90% of POIs are observed less than 10 times.

Twitter has no hierarchic POI categorization in terms of

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

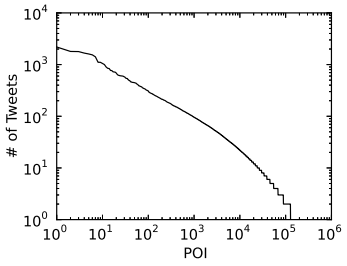

Figure 1: Tweet Distribution over POIs

functionality; Foursquare, on the other hand, does not provide adequate API for crawling data, but offers POI categories organized in a 3-level hierarchy with 9 categories at the top level, such as *food*, *nightlife*, *shops* and 391 subcategories at lower levels, such as *American food*, *bars*, *apparels*. This would enable us to analyze user behavior on different levels of granularity. As no public API exists to map places between the two services, we have to bridge the data from Twitter and Foursquare by using Foursquare's venue search API. This API accepts a query and a pair of coordinates and returns a list of matching places. By carefully mapping places from both sides, we obtain category information (i.e., POI types) for about 85% of the POIs. By grouping tweets into trails as defined in Section 1, and filtering out those trails shorter than 5 visits, we finally have 17 738 trails forming our ground truth. The average length of those trails in terms of visits and places are respectively 7.345 and 5.096. To enable reproduction of our results, we make our anonymous dataset available at `http://homepage.tudelft.nl/9y54n`

3. METHODS AND EXPERIMENTS

Similarly to the work in [2], we assume users' next steps to depend on their current and past locations. Let a trail be a sequence of random variables X_t indicating the categories of POIs, and then simplifying the assumption, we have $P(X_t|X_{t-1}, X_{t-2}, \ldots, X_0) = P(X_t|X_{t-1})$. This fits the definition of Markov Chains. We use them in this exploratory study to verify our assumption of causal dependencies between trail elements. As a baseline, we assume users to stay at their current location type. This is inspired by the observation that users tend to post several times from a single place and its performance is significantly better than random guessing.

The experiments are conducted at two levels of granularity of categories provided by the hierarchy from Foursquare, i.e., the top level categories and sub-level categories. Under both levels of granularity, we repeat the experiment on tweets with POIs located in 4 major US cities, namely, New York (NY), Chicago (CH), Los Angeles (LA) and San Francisco (SF) as well as the whole country, in order to check whether users from the same areas have locally coherent visiting patterns.

The data set in each experiment is split according to 10-fold cross validation. Moreover, our split is based on users, i.e., there is no overlap between training set and testing set in terms of users, which will prevent the methods from overfitting users who have more trails. Due to the sparsity in the data, there are few trails in the test set, e.g., within San Francisco, there are only 200 trails composed of 800 POI-tagged tweets for test. According to the assumption of dependency in the model, we test the method for each consecutive POI pair in each trail. For example, we test (Desserts, Apparel), (Apparel, Record Shop), (Record Shop, Toys&Games) for the trail (Desserts, Apparel, Record Shop, Toys&Games). Due to the setting, there is only *one* true positive in each result list. Accuracy-based measurements would fail distinguishing differences between two rankings having the reference ranked higher than the cut-off. Therefore, Mean reciprocal rank (MRR) is used for evaluation. The results of our experiments are listed in Table 1 and the significance of our results is tested by a Wilcoxon signed rank test ($\alpha < 0.002$).

Table 1: Evaluation in MRR

	Frequency		Subcategory		Category	
	Tweets	Trails	MC	BL	MC	BL
NY	22307	5471	0.400	0.334	0.661	0.618
CH	8313	1835	0.545	0.474	0.750	0.709
LA	9641	2315	0.794	0.777	0.888	0.871
SF	7399	2315	0.591	0.536	0.779	0.733
US	130298	26135	0.469	0.420	0.710	0.668

From Table 1, we can see our trail-causality-based model (MC) outperforming the baseline (BL) significantly in all experiments. On the other hand, the performance of the baseline suggests that there is a strong tendency that users stay at the same POI type for many consecutive postings. By manually checking against the data, we find that there are numerous trails composed of chat with friends posted coherently in both time and space dimensions, i.e., a short burst of tweets from one place. For example, we find one user with two trails full of tweets coming from a single place. The content of those tweets are replies to her friends' tweets. This observation suggests a more thorough study on the time dimension of trails which is left for our future work.

4. CONCLUSION AND FUTURE WORK

We have shown, in our preliminary research, the viability of using Markov Chains to estimate users' next visits in terms of place types based on POI-tagged tweets. By matching places in Twitter and Foursquare, we augment the place information with category information which we would like to share. In the future, we would like to exploit additional sources of evidence such as time of posting or notions of personal preference to further support user tasks that depend on spatial context.

5. REFERENCES

[1] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *WWW '10*, pages 61–70, 2010.

[2] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura. Travel route recommendation using geotags in photo sharing sites. In *CIKM '10*, pages 579–588, 2010.

[3] A. Sadilek, H. Kautz, and J. P. Bigham. Finding Your Friends and Following Them to Where You Are Categories and Subject Descriptors. In *WSDM '12*, pages 723–732, 2012.