Recommending Interesting Events in Real-time with Foursquare Check-ins

Max Sklar
Foursquare
568 Broadway
New York, NY
max@foursquare.com

Blake Shaw
Foursquare
568 Broadway
New York, NY
blake@foursquare.com

Andrew Hogue
Foursquare
568 Broadway
New York, NY
ahogue@foursquare.com

ABSTRACT

Foursquare is a location-based social application that helps users explore the world around them and share their experiences with friends. When foursquare users visit places, they "check in" using their mobile phones, indicating they are at that place. People check in for a variety of reasons: to keep up with friends, get tips about places, redeem rewards, and keep track of their personal history [2]. In aggregate, billions of these check-ins reveal distinct patterns about when places are popular [5] and allow us to build a unique place recommendation engine which can identify and recommend interesting events in real-time based on statistical deviations from past historical trends.

Categories and Subject Descriptors

H.2.8 [**Information Systems Applications**]: Data mining, Spatial databases and GIS; H.3.3 [**Information Systems Applications**]: Information filtering; H.3.5 [**Information Systems Applications**]: Web-based services; G.3 [**Probability and Statistics**]: Time series analysis; I.5.1 [**Pattern Recognition**]: Models—*Statistical*

Keywords

real-time event identification, foursquare, spatiotemporal data, machine learning

1. INTRODUCTION

One of the oldest components of foursquare is a system called "trending" which ranks nearby places by the number of people currently checked in [1] . We use this system to detect interesting "events" where many people check in at the same time. While this is useful, it tends to surface expected events at large venues, such as transportation hubs and office buildings. In order to identify truly interesting events, we have created a system that detects events that are both large and anomalous by leveraging machine learning and past historical data. We call the measure that this system produces "off-trending."

Consider Figure 1, which shows the number of check-ins per hour for two venues over a period of 20 weeks in 2011. Penn Station, shown in Figure 1(a), is one of the most popular places in New York City, and often has more people checked in than any other place in the city. The number of check-ins per hour fluctuates from 0 to well over 200 with a very predictable pattern. There is a strong weekly periodicity with distinct peaks located at commuting times on weekdays. Although these periodic events are quite large in magnitude, they aren't very interesting to users since they appear regularly.

In Figure 1(b), we see a popular New York bar called The Scratcher. This venue has far fewer check-ins per week, and cannot compare to Penn Station in terms of raw popularity. However, there are occasional events at The Scratcher, as shown in week 5 on Friday night, that constitute a large deviation from historical behavior. These are the kinds of events that we aim to identify and surface to our users. Instead of surfacing common popular places like airports and train stations, we aim to surface anomalous events like street fairs, parties, and gallery openings. The key to identifying these events in real-time is modeling the expected number of check-ins each place is likely to receive.

RecSys'12, September 9–13, 2012, Dublin, Ireland, UK.
ACM 978-1-4503-1270-7/12/09.

2. TEMPORAL MODELS OF PLACES

We define the "off-trending" score of a venue, Ω_v, to be an estimate of how *unusually* busy the place is (that is, how busy the venue is relative to historical trends). To calculate the off-trending score, we first learn a probabilistic model for each place that yields a negative binomial distribution [4] over the number of people we expect to check in at any given time:

$$P(k|t,\alpha_w,\beta) = \frac{\Gamma(\alpha_w+k)}{\Gamma(\alpha_w)\Gamma(k+1)}\frac{\beta^{\alpha_w}t^k}{(\beta+t)^{\alpha_w+k}}$$

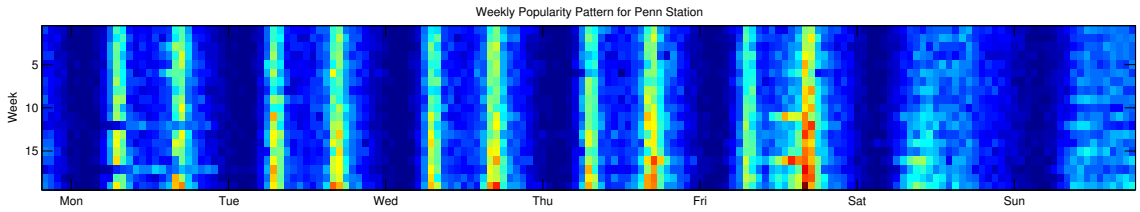
where k is the number of people who have checked in, t is the elapsed time period, and w is the current hour of the week. The parameters of the probabilistic model, α_w, are parameterized by the hour of the week w (a number between 1 and 168), and a tunable parameter β which controls the variance of the distribution:

$$\alpha_w = \beta C_{(10-\text{day})} P(w).$$

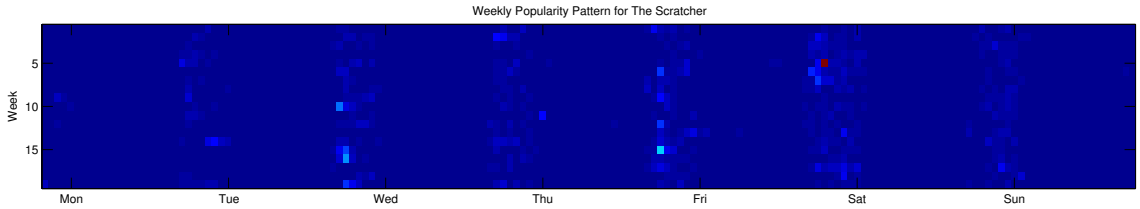
$C_{(10-\text{day})}$ is an estimate of the number of check-ins per week computed using an exponential moving average with a 10-day half life, and $P(w)$ is the probability of a check-in at the place occurring at weekhour w. The term $C_{(10-\text{day})}P(w)$ can be thought of as the expected number of check-ins for a particular hour of the week.

2.1 Learning the Parameters

For each popular place on foursquare we estimate the 168 parameters of $P(w)$ by maximum likelihood from all of the



(a) Penn Station



(b) The Scratcher

Figure 1: Weekly check-in patterns for two places in New York City: Penn Station (top) and The Scratcher (bottom). Each colored cell represents a single hour in a given week. Red entries indicate hours with many check-ins, and blue entries correspond to hours with very few check-ins. Note there are strong periodic events that happen at Penn Station which need to be accounted for in order to not overwhelm more anomalous events at less popular places (such as the event at The Scratcher on Friday of Week 5).

check-ins that have occurred at that place. We use a Dirchlet prior [3] to smooth the estimates and avoid over-fitting:

$$P(w) = \frac{C_w + \theta_w}{\sum_{i=1}^{168}(C_i + \theta_i)}.$$

C_w represents the number of check-ins that occur at the place at week hour w. The hyper-parameters of the Dirchlet distribution, θ_w, are optimized over a large subset of places via gradient descent using the log-likelihood of held-out data as an objective function. Every time a user checks in on foursquare, we increment both the weekhour-based counts C_w, as well as a global popularity term $C_{(10-\text{day})}$, which estimates the number of check-ins a venue receives per hour by averaging check-in counts from many previous days.

2.2 Ranking Places

Given $P_v(k|t, \alpha_w, \beta)$ for each place v in a candidate set with a current value of k people checked in, we can compute $\Omega_v = \sum_{j=0}^{k} P_v(j|t, \alpha_w, \beta)$ based on the cumulative density function of the negative binomial distribution. We call this quantity the "off-trending" score for a venue, which indicates how unlikely it is for us to see at least k people checked in to a place.

2.3 Summary

The off-trending score is a critical component of foursquare's recommendation engine, allowing us to identify and surface interesting nearby events in real-time. When a user opens foursquare's Explore recommendation engine on their mobile device, the off-trending score is combined with signals about the user's past history, the preferences of the user and their friends, and other global signals in order to deliver a ranked list of recommendations for interesting places nearby the user might want to discover. Unlike other recommendation systems such as those for movies or books, understanding timeliness is critical to building a great place recommendation engine that can suggest activities to users in real-time.

3. REFERENCES

[1] foursquare for developers. https://developer.foursquare.com/, 2012.

[2] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman. I'm the mayor of my house: examining why people use foursquare - a social-driven location sharing application. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 2409–2418, New York, NY, USA, 2011. ACM.

[3] T. P. Minka. Estimating a dirichlet distribution. 2000.

[4] T. P. Minka. Estimating a gamma distribution. 2002.

[5] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. In *Proceedings of the 5th Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 570–573, July 2011.