Tip Ranker: A M.L. Approach to Ranking Short Reviews

Enrique Cruz
Foursquare Labs
568 Broadway, 10th Floor
New York, NY
enriquecruz@foursquare.com

Berk Kapicioglu
Foursquare Labs
568 Broadway, 10th Floor
New York, NY
berk@foursquare.com

ABSTRACT

Foursquare is a local search and discovery app where as part of the experience users leave tips, short reviews and suggestions to help other users find great places. This poster summarizes the strategy we use to select the best tips for a given venue. Our new ranking model leverages text, contextual and social signals to selects the tips that provide our users with the most informative and high quality content. The new model has numerous applications within the Foursquare app ecosystem and its introduction yielded significant and positive results in our metrics as measured by various A/B tests.

Categories and Subject Descriptors

I.2.6 [**Learning**]: Machine Learning; I.5.1 [**Pattern Recognition**]: Models–SVM; I.5.4 [**Pattern Recognition**]: Application; H.3.3 [**Information Search and Retrieval**]: Information filtering; I.2.7 [**Natural Language Processing**]: Text analysis

Keywords

machine learning, ranking, svm, context-aware recommenders, Foursquare, A/B test, natural language processing, text classification

1. INTRODUCTION

Foursquare is a location-based recommendation engine. One of the primary actions for our users is to write a tip, which is a short public blurb of text attached to a venue. A tip is often a review or a suggestion about the place. Over the years, Foursquare has collected over 87 million tips from users. This paper describes a system we built that improved upon our previous methods of sorting tips. The new model is designed to rank and select the most relevant and informative tips for our users at a given venue.

2. BACKGROUND

Historically Foursquare has used a few different mechanisms for sorting and selecting the best tip at a venue. None of these were fully satisfactory on their own. We enumerate a few of the most prominent strategies employed and their shortcomings:

RecSys 2016 Poster Proceedings September 15-19, 2016, Boston, MA, USA

Popularity: This is a measure of the positive interactions a tip has garnered since its creation. While generally doing a good job at showcasing content that is relevant or useful to users, it tends to favor content that is old and stale and leads to a feedback cycle where highly ranked tips are more prominently exposed and then become even more popular.

Recency: This is a measure of the amount of time that has passed since the tip was created. It does a good job at keeping the content fresh and showcasing the vibrancy of the Foursquare community, yet it offers no guarantees of quality or relevance.

3. TIP SELECTION AND RANKING

In addition to popularity and recency as defined above, we included the following features in our model:

3.1 Feature Components

Language Identification: A language classifier built using an ensemble of open source and home-grown solutions in order to avoid serving tips in languages that a user does not understand[1].

Content Richness: A number of signals which track more general attributes about the tip, beyond the text information it contains, but that nonetheless affect the way the content of the tip is perceived. Amongst these are the presence or absence of a photo attached to the tip as well as the number of tokens and words the tip contains.

Author Trust: Statistics around the author such as tenure as a Foursquare user as well as total popularity and other aggregate atatistics around the user's previously written tips. These signals represent a user's trusthworthyness as a tip author.

Global Quality: A set of previously built scores from various statistical classifiers that were trained to identify specific traits such as sentiment or spam[1].

3.2 Data and Labels

To collect our data we first determined the top 1,000 most popular venues by user views and proceeded to randomly sample 100 distinct pairs of tips from each of these venues. After accounting for some language filtering and de-duplicating this yielded us a dataset of 50,000 tip pairs.

We proceeded to label this data by designing a CrowdFlower[1] job where the judges would be shown a tip pair from our sample pool alongside the relevant venue. The

[1]A crowdsourcing platform for tasks similar to Amazon Mechanicl Turk

judges then were asked the question "If you where either currently at this venue or considering visiting this venue which of the following pieces of content is more informative?".We designed the job in such a way that the tips would be shown in a context as similar as possible to they way they are displayed in Foursquare, exposing our judges to all the same contextual information that affects the way our users perceive a tip.

The outcome of our CrowdFlower job yielded around 35,000 labeled pairs of tips which we divided into training and evaluation data.

3.3 Training and Evaluation

To train our new tip ranker we explored a variety of algorithms including LambdaMART, Coordinate Ascent and RankBoost. In the end we used SVM$^{\text{rank}}$[2] an implementation of Support Vector Machines as our supervised learning algorithm while our optimization metric was simply trying to minimize the number of misordered pairs of tips with regards to our training labels.

As we iterated and tuned our new ranker, we evaluated its performance against the held out dataset and compared it against some baseline metrics. We also evaluated the rankers heuristically with a new Side by Side tool[2] to look at the best tips for a venue chosen by each model.

Table 1: Tip Ranker Baseline Evaluation

Ranking Function	Percentage Correct
Tip Ranker w. Text Features	87%
Tip Ranker w/o Text Features	84%
Popularity	54%
Recency	51%
Random	50%

In the final model, the features with the highest weight are as follows:

- **Tip Length and number of tokens**
- **Presence of a photo**
- **Positive sentiment**
- **Recency**

The features with the least amount of predictive power turned out to be:

- **Popularity**
- **Author's aggregate statistics**

3.4 A/B Testing and Applications

After the encouraging results of the newly trained tip ranker we brought the model into production to be used on our entire venues corpus and leveraged it into various touchpoints within the Foursquare ecosystem that would benefit from an improved way of selecting and ranking tips. Below we enumerate some of the places we experimented with the new ranker and the results from running A/B test with a 50% split of our userbase.

At a Venue Ping: When we detect that a user is at a given venue with a certain likelihood, Foursquare sends the user a ping containing the best tip (not previously seen by the user) for the venue. This was previously determined using only the global quality features which fed into a random forest model[1] for scoring, sorting as well as filtering tip candidates. Our new ranker yielded significant improvements against the control group, resulting in a 1.5% increase in the CTR while also allowing us send 32% more tip pings by removing some existing hand tuned filters that existed due to a lack of confidence in the prior selection method. Furthermore the experiment group resulted in a 5% increase in core app activity days[3].

Post Check-in Insight: When our users check-in on Swarm[4] we show certain pieces of content for the place the user just checked in. Among these is a Foursquare tip for the venue and an upsell to view all tips if they have the Foursquare app installed or download it otherwise. Previously this tip selection was done purely on social signals. The A/B test with the new model saw a significant increase in all tip related actions (likes, writes, photos) as well as a net lift of 1% active users for Foursquare due to more people choosing the upsell.

Venue Page Default Sort: When displaying a venue page we show a list of the venue's bests tips. This was previously just defaulted to a sort on the positive social signals for the tips. We ran an A/B test sharded on venues in order to measure any SEO changes. While the logged in version of the experiment yielded no significant results the SEO version resulted in a lift of 2.40% in total global referral traffic. We hypothesize that this was mostly driven by the ranker's prefference for content that was longer, included more photos and was written more recently.

4. FUTURE WORK AND EXTENSIONS

There are a few areas of work left to explore that could yield further improvements in the way we select tips by incorporating new features into the model.

Negative Social Signals: At the time the model was built Foursquare provided users only with ways to either like/save a tip or flag it as spam. Since then we have introduced a new interaction to downvote a tip and will retrain the model with this new signal to validate whether it has any predicting improvements.

Sentiment to Rating Matching: The model overwhelmingly prefers tips with positive sentiment. While this is good for a lot of cases, it presents some dissonance when a venue has a low rating yet the top tips are mostly positive. An extension of this work can rank tips to show a sentiment distribution that better reflects the venue's rating and its underlying distribution of votes.

5. REFERENCES

[1] Sklar, M., & Concepcion, K. (2014, September). Timely Tip Selection for Foursquare Recommendations. In Proceedings of the eighth ACM conference on Recommender systems (pp. 1-2). ACM.

[2] Thorsten Joachims (2009, March). Support Vector Machine for Ranking (http://svmlight.joachims.org/). Cornell.

[2]A tool that lets you visualize ranking shifts across a candidate set

[3]Number of days where a user used foursquare and took a core action, such as searching or interacting with a venue

[4]Swarm is Foursquare's companion app, where users can check-in to venues and share their location with friends