

## A Personalized Travel Recommendation System Using Social Media Analysis

Joseph Coelho, Paromita Nitu, and Praveen Madiraju  
Dept. of Mathematics, Statistics & Computer Science  
Marquette University  
Milwaukee, WI, USA.

{joseph.coelho,paromita.nitu,praveen.madiraju}@marquette.edu

**Abstract**—Personalization of recommender systems enables customized services to users. Social media is one resource that aids personalization. This study explores the use of twitter data to personalize travel recommendations. A machine learning classification model is used to identify travel related tweets. The travel tweets are then used to personalize recommendations regarding places of interest for the user. Places of interest are categorized as: historical buildings, museums, parks, and restaurants. To better personalize the model, travel tweets of the user's friends and followers are also mined. Volunteer twitter users were asked to provide their twitter handle as well as rank their travel category preferences in a survey. We evaluated our model by comparing the predictions made by our model with the users choices in the survey. The evaluations show 68% prediction accuracy. The accuracy can be improved with a better travel-tweet training dataset as well as a better travel category identification technique using machine learning. The travel categories can be increased to include items like sports venues, musical events, entertainment, etc. and thereby fine-tune the recommendations. The proposed model lists 'n' places of interest from each category in proportion to the travel category score generated by the model.

**Keywords**—recommender systems; social media analysis; twitter data; personalization

### I. INTRODUCTION

Online services in the virtual world afford more options than a brick and mortar enterprise. When choices are multiplied the act of choosing can become an overwhelming exercise. Recommender Systems (RS) mitigate the problem by filtering the available options to suit the user's tastes and needs. RS have become powerful and popular tools especially in e-commerce operations in this age of information overload. RS has evolved considerably over the last decade[8]. RS filtering can be based either on user-similarity measures (collaborative filtering) or content-similarity measures (content-based filtering). Hybrid RS combine aspects of collaborative and content-based filtering. RS can also be either personalized or non-personalized. Personalization requires some insight into user's likes/dislikes in the given domain. Other improvisations include constraint-based filtering and context-aware filtering.

Travel is one domain that benefits from RS. Planning a vacation can be daunting. Every tourist location has multiple

sights that seem interesting. Websites like TripAdvisor.com and Expedia.com provide information about places based on ratings provided by users of the website. These ratings are not personalized and so may not suit everyone. An automated personalized RS would cater to the needs of the specific user and make planning easier.

A user's social media content is a rich resource for personalizing RS. Smartinsights.com in its latest report states that over 3.1 billion people are active social media users[5]. Over 2.1 billion users are on Facebook, 800 million on Instagram, and 330 million on twitter. Content posted on these platforms can be mined to create user profiles detailing their likes and dislikes. Twitter is a service that allows users to post 280 character-long tweets that may include photos, videos, hash-tags, user mentions, and URLs. Hash-tags are used to indicate tweet topics and help to cluster tweets. Twitter users can be friends with as well as follow other users and this relationship helps in user-collaborative filtering. Tweets generally express opinion. Sentiment analysis helps classify tweets as positive, neutral, or negative. We used twitter data since it is available in the public domain unlike Facebook posts. But twitter does not provide personal user data apart from features like name, location, and language. In many cases users turn the location setting off. Hence in this study we have not taken personal user attributes into consideration.

Martinkus and Madiraju[7] proposed a model that uses twitter data to personalize places of interest (POI) recommendations. Their model takes into account tweets related to travel. User preferences for the different categories are first identified. The tweets are then assigned scores depending on features like favorite count, re-tweet count, and similar user count (a user is assumed to be similar to its followers). These scores are then used to compute the rank value of each category. This paper carries that research forward by including other relevant tweet attributes such as URL count, number of hash-tags, number of user mentions, number of media attachments, length of the tweet, and followers' and friends' preferences that would make for better personalization. In addition, machine learning is used for identifying travel tweets and a bag-of-words model is used for identifying tweets related to specific places of interest categories. The

rest of this paper is organized as follows: Section 2 gives a brief overview of recommendation techniques. Section 3 explains the proposed travel recommendation system. Section 4 includes the conclusion and proposes extensions to this work.

## II. RECOMMENDATION TECHNIQUES

The roots of RS can be traced back to extensive work in cognitive science, approximation theory, information retrieval, forecasting theories and management and marketing science[1]. As a research area it began to gain prominence in the 1990s. Issues like comprehensive understanding of users and items, multidimensionality of recommendations, and non-intrusiveness are aspects of RS that were considered important from early on. Herlocker, Konstan, and Riedl introduce the concept of collaborative filtering and detail the black-box processes involved[6]. Celdrn, et. al. propose a RS based users' behavior and collaborative location and tracking[4]. Bobadilla, et. al. provide a rich survey of RS and discuss two relevant issues: social filtering and content-based filtering[2]. Burke provides a survey of hybrid RS and discusses various hybrid models[3]. The *Recommender Systems Handbook*[8] provides a detailed classification of RS based on techniques used. Accuracy alone is not a measure of a good RS. Diversity needs to be factored in the recommendations so that "niche" objects are not blinded by desire for accuracy[10].

## III. TRAVEL RECOMMENDER SYSTEM

Van Canneyt, et. al. proposed using social media to find places of interest[9]. They showed how geographically annotated social media data could be used to complement existing place databases. But their work does not focus on personalizing the RS using social media. Other works have used photographs posted by users to identify user likes/tastes.

This paper proposes a solution that provides users with recommendations regarding places of interest based on their twitter data. We assume that the user or the user's followers/friends have tweets related to travel. The user is required to provide their twitter handle. The user's (and its followers' and friends') tweets are mined and travel tweets are identified. Travel tweets are then assigned a score based on their features and sentiment. Tweet scores are combined based on POI categories. The solution was developed in Python using the packages *tweepy* (to access twitter data), *BotoMeter* (to identify bots), *TextBlob* (for sentiment analysis), *sklearn* (for machine learning modules), and *Google API* for retrieving places of interest. POI categories considered are Buildings, Museums, Parks, and Restaurants. A user whose tweets have been scored can input a place name and the model will list out POIs in that given place that match the user's likes as predicted by the model.

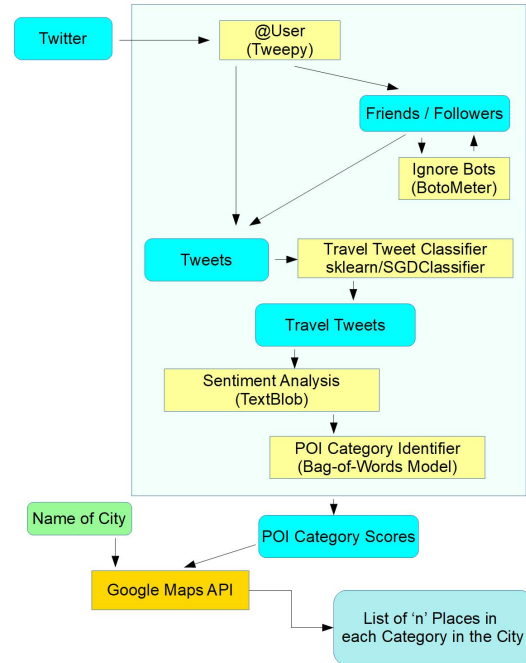


Figure 1. System Architecture

### A. System Architecture

Figure 1 provides a high-level view of the system architecture. A user's tweets have to be collected and classified as "travel" or "non-travel". This can be done either by using a "bag of words" model whereby tweets are scanned for the presence of travel related words like {travel, tour, trip, museum, ...} or by using a supervised machine learning classification model. Travel tweets then need to be classified under the POI categories and this again can be done using bag-of-words or machine learning methods. Sentiment analysis helps identify whether the tweet content is positive, neutral or negative. All these analysis results in tweets getting scores which are then aggregated under POI categories.

The process is repeated for up to ten friends and followers who are "similar" to the user. Similarity is identified if the user has re-tweeted or favorited tweets of the friend or follower, or mentioned the friend or follower in tweets. If ten similar friends and followers are not available the most recently added friends and followers are considered. Friends and followers who might be bots are excluded. The scores of the user, the friends and followers are aggregated under the POI categories to compute the POI category score. These scores are used to determine the actual places of interest to recommend to the user in a given city. To enable diversity, the proposed model displays places of interest in all categories but the number of places in each category will

depend on scores generated for that category.

### B. System Work-Flow

Once the user provides the twitter handle the first step is to collect all the tweets. Twitter allows the collection of up to 3200 most recent tweets of a user. For each tweet a note is made of number of user-mentions, number of URLs, number of media entities, number of hash-tags, favorite count, re-tweet count, and length of the tweet. The tweet is then converted to lowercase and all non-alphanumeric characters are stripped off.

1) *Travel Tweet Classifier*: A supervised machine learning classification model is used to determine whether a tweet is related to travel or not. The model was trained with a dataset of 3000 manually curated tweets that are tagged as travel or non-travel.

2) *User Travel Tweet Scoring*: Each tweet is assigned a score based on the features extracted. The basic score of a tweet is a normalized value given that a tweet can be up to 280 characters long. Tweets having user mentions, hash-tags, URLs, and attached media are given extra points since those features convey more information and add to the importance of the tweet. Favorite count and re-tweet count are also factored in to the score since they too add value to the tweet. Currently the weights for each feature are assigned based on presumed value.

$$\begin{aligned} \text{tweet\_score} = & \text{normalize}(\text{favorite\_count}) + \\ & \text{normalize}(\text{retweet\_count}) + \\ & \text{hashtag\_count} * 0.3 + \\ & \text{url\_count} * 0.4 + \\ & \text{mentions\_count} * 0.3 + \\ & \text{media\_count} * 0.6 + \\ & \text{normalize}(\text{tweet\_length}) \end{aligned} \quad (1)$$

3) *Travel Tweet Categorization*: In this prototype, travel tweets are classified into four categories of interest: historical buildings, museums, parks/outdoors, and restaurants. Words in each tweet are matched with collections of words that identify the category. NLTK's WordNetLemmatizer is used limit the number of words required in the bag-of-words for each category. A tweet is classified based on the category that has the maximum matches.

4) *Sentiment Analysis*: We used TextBlob to perform sentiment analysis and use the polarity score to identify tweets as positive (polarity>0), neutral (polarity=0), or negative (polarity<0). The scores of all tweets for the given category and sentiment are aggregated and then normalized. This provides scores that indicate preferences for the given user  $U_{c,s}$ , where  $c$  is the category,  $s$  refers to sentiment, and  $\max(c, s)$ , the maximum score in the given category is used

to normalize the score.

$$\begin{aligned} U_{c,s} &= \frac{\sum_{c,s} \text{score}}{\max(c, s)} \\ F_{c,s} &= \frac{\sum_{c,s} \text{score}}{\max(c, s)} \\ L_{c,s} &= \frac{\sum_{c,s} \text{score}}{\max(c, s)} \end{aligned} \quad (2)$$

5) *Extending the Model*: The second stage of the user-user collaboration is identifying up to ten friends (following) of the given user who might have similar interests. Only friends with at least 15 travel tweets are considered. The scores of the ten friends are aggregated based on category and sentiment and then normalized using the same process as before to obtain  $F_{c,s}$ . The third stage repeats the same process for up to ten followers. They are mined as before and scores are obtained for the given categories and sentiments  $L_{c,s}$ .

6) *The Final Model*: The positive, neutral and negative scores of the user, the friends, and the followers are separately aggregated for each category. However, the scores for the user, friends, and followers cannot be considered on par. Different weights need to be assigned for each of them, indicating their level of importance, while computing the final scores for each category.

*Calculating Weights*: From the survey conducted we got the actual user preference scores of 22 users for each category ( $S_b, S_m, S_p, S_r$ ) and from the model we have tweet scores and categories for each tweet. We obtained the coefficients ( $\beta_b, \beta_m, \beta_p, \beta_r$ ) for each of the categories using logistic regression. The highest  $\beta$  for each category was summed across all users to obtain  $\beta_U$  which is the weight to be assigned to the users' tweet scores. These weights would be updated periodically to reflect the changing data.

This process is repeated with all the tweets of the user's friends to obtain  $\beta_F$  which is the weight assigned to the friends' tweet scores; and with the tweets of the user's followers to obtain  $\beta_L$  which is the weight assigned to the followers' tweet scores.  $\beta_U, \beta_F$ , and  $\beta_L$  are normalized such that their total equals 10. The weights we obtained using the above method were  $\beta_U = 3$ ,  $\beta_F = 4$ , and  $\beta_L = 3$ .

The tweet scores for each category for the user, the friends, and the followers is aggregated based on sentiment. The scores for positive sentiment have a weight of 1.0, neutral sentiment have weight 0.65, and negative sentiment have weight 0.35. The final scores are sorted to obtain the preferred list of categories.

$$\begin{aligned} U_c = & U_{c+} * \beta_U + F_{c+} * \beta_F + L_{c+} * \beta_L + \\ & (U_{cn} * \beta_U + F_{cn} * \beta_F + L_{cn} * \beta_L) * 0.65 + \\ & (U_{c-} * \beta_U + F_{c-} * \beta_F + L_{c-} * \beta_L) * 0.31 \end{aligned} \quad (3)$$

In the above equation  $c$  refers to the categories while  $+$ ,  $n$ ,  $-$  refer to the sentiment.  $U$  refers to the user,  $F$  to the friends, and  $L$  to the followers.  $U_c$  refers to the final score for category  $c$  for the user  $U$ .

Table I  
CATEGORY SCORES.

Category	Score
Parks	0.37
Museums	0.28
Restaurants	0.21
Historical Buildings	0.14

The above section details the user-user collaboration in identifying friends and followers who share similar interests with the given user and draws up a ranked score for the specified categories. The result will be a list of scores as shown in the following example result for a particular user (Table 1). The prototype developed provides up to fifteen recommendations to the user and the recommendations are proportionately drawn based on the scores computed. Thus for the above example there will be 6 POIs for parks, 4 for museums, 3 for restaurants, and 2 for historical buildings in the recommendation list provided to the user.

The predictions made by the model were matched against the user preferences given in the survey. The overall accuracy of the predictions across all four categories got was 62%. After observing that most tweets came from 'friends' and 'followers' rather than from the users themselves we proposed to set  $\beta_U = 1$ ,  $\beta_F = 6$ , and  $\beta_L = 3$ . This resulted in an accuracy of 68%.<sup>1</sup>

One of the reasons for low accuracy could be the failure to adequately categorize travel tweets. At present tweets are categorized based on words contained in the tweet. This could be improved by using machine learning to identify categories instead of a static classification method based on a set of words. Further, the model's inability to correctly identify categories could be due to the lack of category specific tweets in the training dataset.

#### IV. CONCLUSION

Recommender Systems (RS) are an essential tool in the age of information overflow. Non-personalized RS are useful in some contexts but personalizing recommendations add value. Social media provides a platform for mining data that can be used to make personalizations since users post opinions on various topics. This project mines data from twitter to personalize travel recommendations.

The proposed model takes into account many tweet features like URL count, hash-tag count, favorite count, etc. that lend value to a tweet. This information can be used to separate general tweets from tweets that might be more valuable. In addition links between users and their friends and followers also provide useful information about the user. All these are factored in the proposed model.

<sup>1</sup>A detailed evaluation could not be included due to lack of space. A longer version of this paper is available at <https://www.overleaf.com/read/srthbsfpcmv>

The overall accuracy of the model is 68% and further work may enhance the accuracy. Curating more travel tweets for the training dataset would enhance the ability of the model to classify travel tweets. Identifying categories of places could be transformed using machine learning to categorize the tweets. The model can also be improved by including reinforced learning whereby user feedback is obtained to fine-tune the model and enhance its predictive ability.

While this paper deals with only four categories of places, additional categories could be added. These could include sites specific to children, young adults, seniors, students on educational tours, etc. The scope of this project could be expanded by mining data from additional sources like LinkedIn or Facebook to complement the user profile generated by data from twitter.

#### REFERENCES

- [1] G. Adomavicius and A. Tuzhilin, *Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions*, IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734-749, Jun. 2005.
- [2] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutierrez, *Recommender systems survey*, Knowledge-Based Systems, vol. 46, pp. 109-132, Jul. 2013.
- [3] R. Burke, *Hybrid Recommender Systems: Survey and Experiments*, User Modeling and User-Adapted Interaction, vol. 12, no. 4, pp. 331-370, Nov. 2002.
- [4] A. H. Celdrn, M. G. Prez, F. J. Garca Clemente, and G. M. Prez, *Design of a recommender system based on users' behavior and collaborative location and tracking*, Journal of Computational Science, vol. 12, pp. 83-94, Jan. 2016.
- [5] *Global Social Media Research Summary 2018*. Smart Insights, March 28, 2018. <https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/> accessed on 30 April 2018.
- [6] J. L. Herlocker, J. A. Konstan, and J. Riedl, *Explaining Collaborative Filtering Recommendations*, in Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, New York, NY, USA, 2000, pp. 241-250.
- [7] P. Martinkus and P. Madiraju, *Personalizing Places of Interest Using Social Media Analysis*, Milwaukee, WI, 2015.
- [8] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds., *Recommender Systems Handbook*. Boston, MA: Springer US, 2011.
- [9] S. Van Canneyt, O. Van Laere, S. Schockaert, and B. Dhoedt, *Using Social Media to Find Places of Interest: A Case Study*, in Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, New York, NY, USA, 2012, pp. 2-8.
- [10] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang, *Solving the apparent diversity-accuracy dilemma of recommender systems*, PNAS, vol. 107, no. 10, pp. 4511-4515, Mar. 2010.