# New Yorker Melange: Interactive Brew of Personalized Venue Recommendations

Jan Zahálka
ISLA, Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
j.zahalka@uva.nl

Stevan Rudinac
ISLA, Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
s.rudinac@uva.nl

Marcel Worring
ISLA, Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
m.worring@uva.nl

## ABSTRACT

In this paper we propose *New Yorker Melange*, an interactive city explorer, which navigates New York venues through the eyes of New Yorkers having a similar taste to the interacting user. To gain insight into New Yorkers' preferences and properties of the venues, a dataset of more than a million venue images and associated annotations has been collected from Foursquare, Picasa, and Flickr. As visual and text features, we use semantic concepts extracted by a convolutional deep net and latent Dirichlet allocation topics. To identify different aspects of the venues and topics of interest to the users, we further cluster images associated with them. New Yorker Melange uses an interactive map interface and learns the interacting user's taste using linear SVM. The SVM model is used to navigate the interacting user's exploration further towards similar users. Experimental evaluation demonstrates that our proposed approach is effective in producing relevant results and that both visual and text modalities contribute to the overall system performance.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.5 [**Information Storage and Retrieval**]: Online Information Services

## General Terms

Algorithms; Human Factors; Experimentation

## Keywords

Interactive city exploration; social media; user-centered design; deep nets; semantic concept detectors; topic models

## 1. INTRODUCTION

New York's reputation of being a melting pot of cultures is well-earned. Millions of people with diverse ethnicities, nationalities, and tastes live in the city, contributing to its vibrant, cosmopolitan atmosphere. Tens of millions of visitors come to the Big Apple yearly to enjoy the city, pondering what places in New York to visit. There are numerous resources that help with the choice, such as Foursquare, Yelp, or TripAdvisor. These are excellent when searching venues enjoying general popularity. Does *vox populi*, however, sufficiently reflect the individuality of New Yorkers?

Personal prism can reveal interesting places off the beaten track. Indeed, virtually everyone has "this little place" they like to frequent in their hometown. These places often do not show up in conventional travel guides and recommender systems. However, if the patrons of these venues exhibit a taste similar to our own in their online multimedia trail, we might want to visit them. Hence, we propose the *New Yorker Melange*, an interactive venue exploration system which allows the interacting users (hereafter referred to as *actors*) to brew their own melange of venues frequented by New Yorkers (hereafter referred to as *users*) similar to them.

Understanding user's preference towards a particular venue and the aspects that make two venues similar is a non-trivial task. Although venue categories (e.g., art museum, bar, park, and restaurant) assigned by some social media platforms make the analysis easier, in practice the variety of aspects of interest to the users may be considered too wide to be captured by such categories. We conjecture that those aspects may be naturally encoded in venue-related images and the metadata associated with them. While some location-based social networking platforms, such as Foursquare, offer a detailed venue description together with the images of the venues contributed by the users, with regard to content quantities and topical richness of community-contributed metadata, they can hardly match content-sharing platforms (e.g., Flickr and Picasa). Therefore, in our approach we choose to utilize a large amount of information automatically collected from multiple platforms, which we effectively integrate by performing topical analysis in visual and text domain.

In Section 2 we reflect on the related work. In sections 3 and 4 we describe our data collection procedure and our proposed interactive venue exploration system. The proposed approach is evaluated in Section 5, while Section 6 provides concluding remarks.

## 2. RELATED WORK

In recent years, a number of location recommendation and exploring applications using community-contributed content have been proposed. For example, Pang et al. mine user-generated travelogues to select representative Flickr images for a particular tourist destination [7]. Kofler et al. con-

*MM'14,* November 3–7, 2014, Orlando, Florida, USA.
Copyright 2014 ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2656403.

ceived a system analyzing users' previously-captured Flickr images to provide personalized recommendations of off-the-beaten-track locations at a destination city of interest [5]. Popescu et al. analyzed a collection of Flickr images to discover existing tourist routes within a city and recommend new ones [8]. Another personalized travel recommendation system, proposed by Cheng et al., recommends optimal routes for a particular user demographic based on user attributes extracted from community-contributed images [2]. Rudinac et al. proposed a visual summarization approach, using community-contributed images and associated metadata to discover various aspects of a geographic area [9]. Finally, Zhao et al. devised a multimodal approach to detecting overlapping communities in Foursquare [12].

What sets us apart from related work is an interactive approach to venue recommendation and city exploring in which the user preferences are learned on-line. Additionally, in our approach the venues are recommended indirectly, as a part of the like-minded users' profiles, which puts a strong accent on user-centricity. Finally, to gain a better insight into user preferences and venue properties, we integrate multimedia data from content-sharing and location-based social networking platforms.

## 3. DATA COLLECTION

The first step of the data obtaining process involved getting the list of venues of interest. To this end, we queried the Foursquare API using geo coordinates corresponding to New York City. This resulted in a JSON database containing the respective venue listings. Each listing comprises information about the venue, such as *name*, *venue category*, and *geo coordinates*. In total, we obtained information about 7,246 verified venues. For each venue, we downloaded up to 200 images associated with it, resulting in a collection of 429,921 Foursquare images.

To diversify, enrich, and enlarge our dataset, we have systematically crawled Picasa and Flickr, two popular image sharing social sites. Each query targeted a particular venue, using the venue name and geo coordinates as parameters. To alleviate incorrect geo coordinate assignments by camera GPS systems and users, we set the geo radius, i.e., the tolerated deviation from the venue's geo coordinates, to 1 kilometer. Each query retrieved up to 500 images along with associated annotations. This process yielded 243,292 images from Flickr, and 398,968 images from Picasa.

In total, our collection contains 1,072,181 images of New York and associated annotations. We believe we have significantly increased the signal-to-noise ratio in our dataset by including both the venue name and the geo coordinates in the query. The quality of the dataset can be further improved at the expense of the size by reducing the geo radius and/or reducing the number of images retrieved per venue. Furthermore, grouping images by categories of interest opens up exciting possibilities for further research. Moreover, the same type of data can be obtained for any location in the world. Our dataset thus presents an interesting data resource for multimedia research in general.

## 4. APPROACH

This section discusses our approach to providing content-driven topical recommendations of venues frequented by New Yorkers. The data preprocessing pipeline, conceptually depicted in Figure 1, consists of three key steps: extraction of visual features from the images (Section 4.1), extraction of text features from the image annotations (Section 4.2), and representing individual venues and users as topical clusters (Section 4.3). Finally, Section 4.4 describes the interactive exploration system as used by the actor.

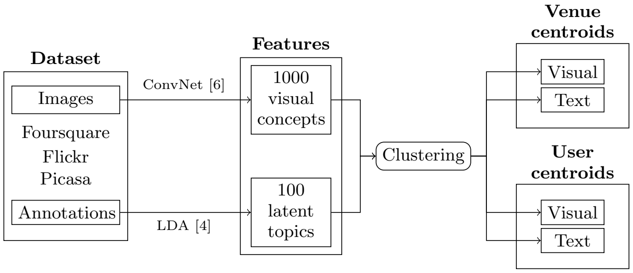

**Figure 1: Data collection and preprocessing.**

### 4.1 Visual Features

Guided by an assumption that the venues of interest to the actor may share visual attributes, we extract visual features from the images associated with them. To this end, we use a deep convolutional neural network as conceived by Krizhevsky et al. [6]. The deep net was trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 dataset of 1000 semantic concepts [3]. We choose to represent each Foursquare, Flickr, and Picasa image by the respective output of the deep net, i.e., the distribution over the 1000 semantic concepts.

### 4.2 Text Features

For all Flickr images we index the text from *title*, *description*, and *tags*. In case of Picasa images, we utilize the *title*, *summary*, *description*, *keywords*, *albumtitle*, *albumdesc*, and *snippet* fields. We further tokenize the text using the Natural Language Toolkit (NLTK) [1] and remove all HTML elements. Finally, we remove common stopwords and words appearing only once. To index the text in a persistent manner, we use the Gensim framework [11] designed for topic modeling of large corpora. After computing the bag of words representation for each document, we perform online Latent Dirichlet Allocation (LDA) [4] to identify a small number of topics associated with the images. To ensure a reasonable granularity, we set the number of topics to 100.

### 4.3 Clustering

Processing million images places a non-trivial computational load on the machine, thus prohibiting interactivity. Hence, a compact representation of individual venues and users is needed. We conjecture that each venue and user has a number of topics of interest within the respective set of images. For example, images of a restaurant might depict the food, the interior, or groups of people enjoying their drink. A user's images, while potentially diverse and numerous, will typically share topics corresponding to the user's main interests. Determining a reasonable number of topics to represent users and venues provides not only a more compact representation, but also a basis for matching them.

To obtain the topics of interest, we cluster all images associated with a venue using the $k$-means clustering algorithm. The cluster centroids represent the individual topics. Simi-

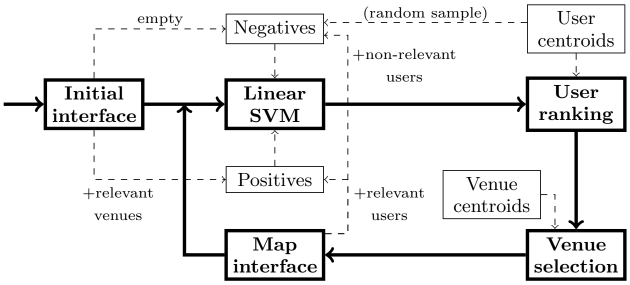

**Figure 2: Interactive exploration; thick components represent the system flow, non-thick nodes and dashed arrows depict the data flow.**

larly, we apply the $k$-means clustering algorithm to cluster all images uploaded by a particular user and select cluster centroids to represent user interests. To ensure a reasonable topical granularity, we choose to use $k = 5$ clusters. The clustering is performed independently for the visual and the text domain.

## 4.4 Interactive Exploration

The interactive exploration pipeline within the New Yorker Melange, conceptually depicted in Figure 2, can be decomposed into five steps described in this section. Steps 2–4 take around a second on a standard PC, which makes New Yorker Melange a responsive, truly interactive system.

**Step 1: Initial interface**: The actor is first presented with a set of images arranged in a grid interface. These images illustrate individual venues, and each is the top-ranked Foursquare image for the respective venue. The actor selects the venues of interest by clicking on the images. Then, two sets of positive examples are initialized to contain the cluster centroids of the clicked-on venues for each respective modality (cf. Section 4.3). The sets of negative examples (again, one for each modality) are initially empty.

**Step 2: Linear SVM**: To identify users similar to the actor, we train a linear SVM separately for visual and text modalities using the respective sets of positive and negative examples as training data. If the size of the set of negatives for any modality is smaller than two times the size of positives, the system supplements the set of negatives with random samples from the collection of user centroids to ensure good positives-negatives balance.

**Step 3: User ranking**: Each user's relevance score per modality is the maximum of SVM scores assigned to the user's cluster centroids, i.e., each user is as relevant as their most relevant topic. The overall user relevance ranking is determined by aggregating the score rankings across modalities using Borda count. The top-5 ranked users are selected to be displayed to the actor.

**Step 4: Venue selection**: For each displayed user, up to 5 relevant visited venues are selected. The venue representativeness score per modality is the maximum of cosine similarities between the venue's cluster centroids and the user's cluster centroids. These scores are again aggregated using Borda count.

**Step 5: Map interface**: The users and their venues are displayed within the map interface depicted in Figure 3. Each venue is represented by a thumbnail of the same image as in step 1. Each thumbnail is placed on the map of

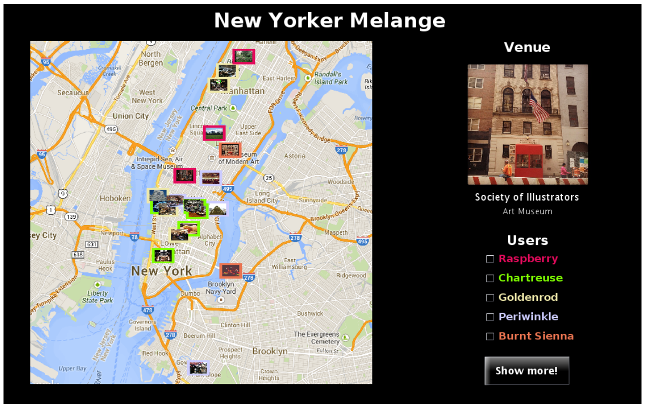

**Figure 3: A screenshot with user recommendations generated by our New Yorker Melange system.**

New York at the venue's coordinates and color-coded by the visiting user(s). The users themselves are anonymized. The anonymization does not prohibit the manifestation of individual New Yorker stories, but ensures the users' privacy. Clicking on an image thumbnail enables the actor to inspect the image and basic information about the venue. The actor can indicate which user(s) are relevant using the checkboxes. Upon pressing the "Show more!" button, the cluster centroids of users indicated as relevant are added to the positive data and the cluster centroids of users not indicated as relevant are added to the negative data. The cluster centroids of all five displayed users are removed from the collection of user centroids, and these users will not show up in subsequent iterations. After these adjustments to the data, the system proceeds with step 2.

# 5. EXPERIMENTAL RESULTS

To evaluate the effectiveness of our system in recommending the users and venues of interest to the actor, we apply a protocol inspired by the common practices in evaluation of recommender systems [10]. First, we select 100 users with the largest number of visited venues to be used as artificial actors. For each, we use 4-fold cross-validation as follows. In each step, 25% of the venues visited by the user are withheld as test data. The remaining venues are indicated as relevant (cf. Section 4.4, step 1). The artificial actor then goes through 10 interaction rounds (cf. Section 4.4, steps 2–4). Whenever it encounters a venue from the test set, it marks the corresponding user(s) as relevant. The system's performance at each interaction round is measured by recall, i.e., the ratio of the number of encountered test venues to the total number of test venues, averaged across four folds.

As the *baseline* for comparison we use an interactive recommender system architecture, utilizing the *user-venue matrix*, which encodes users' preference towards venues. The values in the matrix are 1 when the user visited a particular venue, and 0 otherwise. Initially, the actor is represented with a binary vector of relevant venues. To identify users similar to the actor (cf. Section 4.4, step 2), we compute the Jaccard similarity coefficient between the actor's vector and the vectors of all users in the collection. The five users most similar to the actor are then selected and displayed on the screen. The displayed users are represented by their visited venues, selected and sorted again by Jaccard similarity to

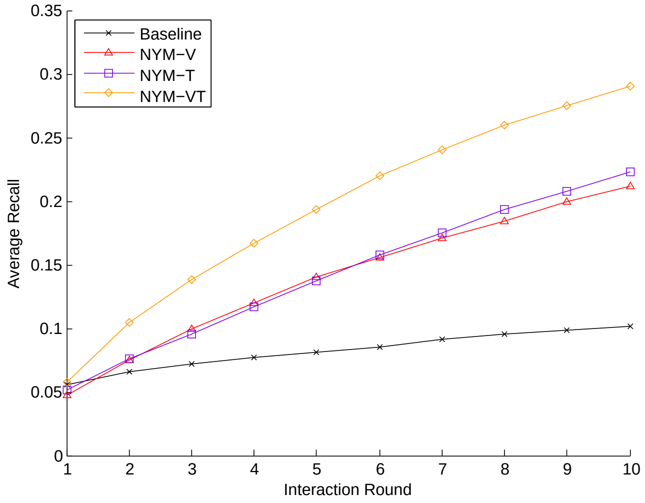

**Figure 4: Performance comparison of several variants of our New Yorker Melange system; the performance is expressed in terms of average recall at interaction round.**

the actor's profile. When the actor selects one of the displayed users, all venues visited by that user are added to the actor's venue vector.

The results presented in Figure 4 show that the New Yorker Melange (NYM-VT) clearly outperforms baseline at each interaction round. This proves the added benefit of topical modeling of venues and users. While the recall of 0.29 yielded by our approach at round 10 may not seem particularly high at a first glance, one should not forget that the evaluation protocol is rather challenging. Note that the artificial actors will not mark displayed users that prefer different venues of the same type as relevant. A closer analysis of the experimental results reveals that when the set of the actor's categories of interest is reasonably sized ($\sim$ 5 categories), the system yields a performance well above average. Conversely, the cases where the performance is lacking are those when the system tries to learn too many categories at once. This is an expected result because, intuitively, the taste of users having preference towards many different categories is difficult to model. This apparent weakness can be easily overcome by splitting the exploration of many different topics into several smaller sessions, each focusing on a small number of topics at once.

We further investigate contribution of each modality to the overall performance of our proposed approach. Therefore, in Figure 4 we compare the performance of the New Yorker Melange system (NYM-VT) with its variants utilizing only visual (NYM-V) and text (NYM-T) information. The results show that both visual and text modalities contribute equally to the overall system performance and that combining them leads to a clear performance improvement.

## 6. CONCLUSION

We have presented the New Yorker Melange, a system that gives the users an opportunity to interactively explore the city through the eyes of New Yorkers sharing their taste. Our experiments have shown that the proposed user-centric approach, which treats venues as part of the like-minded user profiles and iteratively learns topical preferences of the interacting user, is effective in producing relevant recommendations. The experiments further suggest that even though useful information about venue properties and user preferences may be extracted from both visual and text content, the performance of our multi-modal approach cannot be matched by its uni-modal variants. Additionally, our work led to a good-quality, large-scale and cross-platform dataset whose value extends beyond the use case addressed in this paper. By performing topical modeling in visual and text domain on the dataset, our approach succeeds in integrating communities and content originating from very different social media platforms.

## 7. ACKNOWLEDGMENTS

This research is partly supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO), and which is partly funded by the Ministry of Economic Affairs.

## 8. REFERENCES

[1] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python.* " O'Reilly Media, Inc.", 2009.

[2] A.-J. Cheng, Y.-Y. Chen, Y.-T. Huang, W. H. Hsu, and H.-Y. M. Liao. Personalized travel recommendation by mining people attributes from community-contributed photos. In *ACM MM*, pages 83–92, 2011.

[3] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and F.-F. Li. Large scale visual recognition challenge 2012. `www.image-net.org/challenges/LSVRC/2012`.

[4] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In *NIPS*, pages 856–864, 2010.

[5] C. Kofler, L. Caballero, M. Menendez, V. Occhialini, and M. Larson. Near2me: An authentic and personalized social media-based recommender for travel destinations. In *ACM WSM*, pages 47–52, 2011.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105. 2012.

[7] Y. Pang, Q. Hao, Y. Yuan, T. Hu, R. Cai, and L. Zhang. Summarizing tourist destinations by mining user-generated travelogues and photos. *Comput. Vis. Image Und.*, 115(3):352 – 363, 2011.

[8] A. Popescu, G. Grefenstette, and P.-A. Moëllic. Mining tourist information from user-supplied collections. In *ACM CIKM*, pages 1713–1716, 2009.

[9] S. Rudinac, A. Hanjalic, and M. Larson. Generating visual summaries of geographic areas using community-contributed images. *IEEE TMM*, 15(4):921–932, 2013.

[10] G. Shani and A. Gunawardana. Evaluating recommendation systems. In *Recommender Systems Handbook*, pages 257–297. Springer US, 2011.

[11] R. Řehůřek and P. Sojka. Software framework for topic modelling with large corpora. In *LREC*, pages 45–50, 2010.

[12] Y.-L. Zhao, Q. Chen, S. Yan, T.-S. Chua, and D. Zhang. Detecting profilable and overlapping communities with user-generated multimedia contents in LBSNs. *ACM TOMCCAP*, 10(1):3:1–3:22, 2013.