# Using Flickr Geotags to Predict User Travel Behaviour

Maarten Clements†, Pavel Serdyukov†, Arjen P. de Vries‡ and Marcel J.T. Reinders†
†Delft University of Technology, The Netherlands. m.clements,p.serdyukov,m.j.t.reinders}@tudelft.nl
‡CWI, The Netherlands. Arjen@acm.org

## ABSTRACT

We propose a method to predict a user's favourite locations in a city, based on his Flickr geotags in other cities. We define a similarity between the geotag distributions of two users based on a Gaussian kernel convolution. The geotags of the most similar users are then combined to rerank the popular locations in the target city personalised for this user.

We show that this method can give personalised travel recommendations for users with a clear preference for a specific type of landmark.

**Categories and Subject Descriptors:** H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

**General Terms:** Algorithms, Experimentation

## 1. PERSONALISED TRAVEL GUIDES

Before visiting a city, many people consult a travel guide or website that lists the most interesting locations. These travel guides are commonly based on the opinions of all other users. However, people have different preferences and therefore are not equally satisfied by these popularity rankings.

We propose to predict a user's favourite locations in a city based on his travel behaviour in previously visited cities. On social photo sharing websites like `www.flickr.com` people can annotate their photos, including the geographical location where the photo was made. Also, increasingly more cameras and smartphones are automatically storing the GPS coordinates when a photo is made. These *geotags* give an accurate indication of the user's preferred landmarks. Based on a set of collected geotags, we define a measure to identify similar users in previously visited cities. Then we aggregate these users' opinions in a different city to obtain a personalized travel recommendation for the target user.

The exploitation of geotags has shown to be effective for various tasks, like global event detection [3] and mapping textual tags to geographical locations [1]. Based on users' GPS tracks, location recommenders have been proposed that attempt to predict popular places and activities near the current location of the user [5, 4].

In this work we predict relevant locations based on users' geotags in a geographically remote location. We show statistical improvements over all users that visited the 10 largest cities and give an effective recommendation example based on an artificial user profile.


*SIGIR'10,* July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

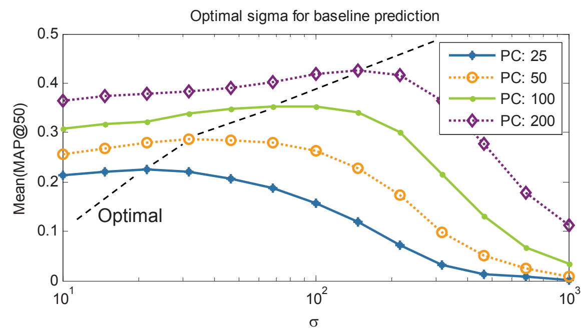


**Figure 1: Mean MAP@50 of the baseline prediction in the top-10 cities for several positive cutoff values ($PC$) and increasing values of the kernel size ($\sigma$).**

## 2. FLICKR GEOTAGS

Using the public Flickr API we have collected the geotags of 36,264 users, who actively use the geotag functionality. Together these users have uploaded 52,425,279 photos of which 22,710,496 have been geotagged.

We keep the data points that lie within the bounding boxes of the ten most visited cities. Based on our data, these cities in order of visitors are: *London*, *New York*, *Paris*, *San Fransisco*, *Los Angeles*, *Rome*, *Chicago*, *Washington*, *Barcelona*, *Berlin*. We only keep users who have made at least 5 photos in at least two cities. After this constraint the number of geotags of a single user in a city ranges from 5 to 5073 photos, and in total the 4750 remaining users have made 12,669 city visits. Together the users made 526,827 photos on the qualifying trips.

## 3. METHODOLOGY & RESULTS

**Baseline ranking** As a baseline prediction we create a scale space representation of all the geotags in each city using a mean shift algorithm, similar to Crandall et al. [1], but using a Gaussian kernel instead of a uniform disc: $K(z) = e^{-z^2/2\sigma^2}$, where the standard deviation $\sigma$ is used as a scaling parameter. This method finds the maximum values of a kernel convolution of the distribution of all users' geotags with a Gaussian kernel ($\Phi_{All}$). To ensure we reach all local maxima, we initiate the mean shift algorithm with all individual geotags. For each subsequent scale we use the peaks found in the previous scale to initiate the optimalisation procedure. The ranking based on the resulting peak

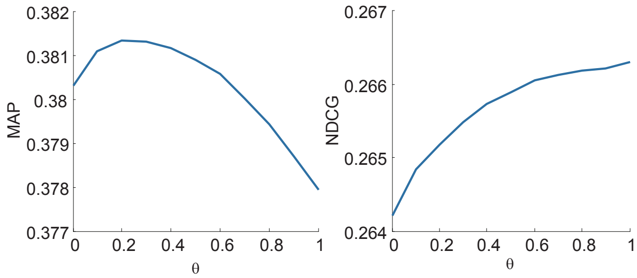


**Figure 2: MAP@50 and NDCG@50 for increasing personalisation weight $\theta$.**

weights gives us the top landmarks for each city, based on the general popularity.

To evaluate the ranking we judge a recommended location $l_j$ as correct if the target user $u_t$ has a geotag $i$ within the positive cutoff value ($PC$) of that location, $\exists i : |l_j - u_t(i)| < PC$. Figure 1 gives the mean MAP@50 over the 10 cities, which computes the mean over the precision after each correct prediction in the top-50.

Figure 1 shows that the optimal $\sigma$ is strongly dependent on the choice of $PC$. The predictions in this paper will be evaluated at $PC = 100$ meter, which is roughly the radius of a landmark (e.g. the Colosseum is $189m$. long). Based on the baseline results at $PC = 100$ we select $\sigma = 68m$. for all further experiments.

**Personalised reranking** To personalise the landmark ranking for $u_t$ in the target city ($C_t$), we compute the similarity between $u_t$ and all other users $u_c$ in the similarity city ($C_s$), where $C_t$ and $C_s$ are any two cities from the top-10, both visited by $u_t$. Using the mean shift algorithm we compute the peaks of $u_t$ at $\sigma = 68m$ in $C_s$. For each peak $k$ of $u_t$ we now compute the value of the kernel convolution ($\Phi_{u_c}(k)$) on the geotags of $u_c$ in $C_s$. The similarity between the two users is now derived by computing the sum over the minimum value in the two resulting profiles $S(u_t, u_c) = \sum_k min(\Phi_{u_t}(k), \Phi_{u_c}(k))$. As both profiles are normalised, this will give a similarity score in the range 0-1.

Based on all similar users we now rerank the top-50 popular locations $l_j$, predicted by the baseline method in $C_t$. This is done by recomputing the kernel convolution at these locations while weighing each user's geotags with his similarity to the target user: $\Phi_{Sim}(l_j) = \sum_{u_c} S(u_t, u_c)\Phi_{u_c}(l_j)$. The top-50 locations are now reranked by a linear combination of the baseline and the personalised score: $R(l_j) = (1-\theta)\Phi_{All}(l_j) + \theta\Phi_{Sim}(l_j)$.

Figure 2 gives the mean results over all users in the 90 possible combinations of two cities. The baseline is represented by the score at $\theta = 0$, where all user similarities are set to 1. Compared to the baseline, the optimal result on MAP improves 0.3%. At $\theta = 0.2$ there are 10,081 trips where we present an improved ranking to the user, against 8,440 trips where the baseline ranking would have been better. We also show the NDCG (refer to [2] for details) where the gain of each correct prediction is assigned as the inverse popularity of that location ($1/\Phi_{All}(l_j)$). The increase in NDCG shows that our recommender suggests less popular and therefore more serendipitous locations.

The improvement on MAP@50 is statistically significant in 22 out of 90 city pairs (based on a paired t-test with $p < 0.05$). For most users the improvement will however not make a big practical difference in the recommended locations. Compared to traditional collaborative filtering data sets, we find that many more people conform to the global popularity ranking if landmarks are concerned. For example, almost all people who visit Paris will make a photo of the Eiffel tower, while people who do not like Sci-Fi movies will never watch *Star Wars* even though it is one of the most highly ranked movies all times. This makes improving over the baseline a challenging task. Also, we observe many mixed preferences in user profiles (e.g. there are no users who *only* make photos at zoos), this makes it hard to match similar users.

**Table 1: Query: *MACBA + Miro***

| City | Rank | ΔRank | Landmark name |
|---|---|---|---|
| London | 1 | +3 | Tate Modern |
| NY | 1 | +10 | Guggenheim Museum |
| NY | 3 | +5 | Museum of modern art |
| Paris | 3 | +4 | Centre Pompidou |
| SF | 3 | +7 | SF Museum of Modern Art |
| Chicago | 2 | +29 | Museum of Contemporary Art |
| Washington | 1 | +20 | Hirshhorn Museum |
| Berlin | 7 | +40 | Hamburger Bahnhof Museum |
| Berlin | 14 | +16 | Neue Nationalgalerie |

As an example of the potential benefit of personalized travel recommendations, we created an artificial user profile with 10 geotags scattered around two modern/contemporary art landmarks in Barcelona (MACBA and Miro foundation). Table 1 shows a completely personalised ranking (with $\theta = 1$) and the rank difference between the baseline and the personalised ranking for modern art museums in other cities. It is clear that in all other cities where a modern art museum was in the top-50 we obtain a big rank improvement between the baseline and the predicted ranking.

**Conclusions** A user's favourite landmarks in a previously unvisited city can be predicted by reranking the most popular locations based on users with similar travel preference. Our results indicate that statistical improvement over all users is hard to achieve, but for users with a clear travel preference very accurate predictions can be made.

# 4. REFERENCES


[1] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW '09: Proceeding of the 18th international conference on World Wide Web*, pages 761–770, 2009.

[2] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[3] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *SIRIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110, New York, NY, USA, 2007. ACM Press.

[4] Y. Takeuchi and M. Sugimoto. Cityvoyager: An outdoor recommendation system based on user location history. In J. Ma, H. Jin, L. T. Yang, and J. J. P. Tsai, editors, *Ubiquitous Intelligence and Computing*, volume 4159, chapter 64, pages 625–636. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[5] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with gps history data. In *WWW '10: Proceeding of the 19th international conference on World Wide Web*, page 10, New York, NY, USA, April 2010. ACM.