Finding similar destinations with Flickr geotags

Davide Eynard
University of Lugano
ITC - Institute for
Communication Technologies
Via Buffi, 13
6900, Lugano - Switzerland
davide.eynard@usi.ch

Alessandro Inversini
University of Lugano
ITC - Institute for
Communication Technologies
Via Buffi, 13
6900, Lugano - Switzerland
alessandro.inversini@usi.ch

Leonardo Gentile
Politecnico di Milano
Via Ponzio 34/5
20133, Milano - Italy
leonard.gentile@gmail.com

ABSTRACT

The amount of geo-referenced information on the Web is increasing thanks to the large availability of location-aware mobile devices and map interfaces. In particular, in photo collections like Flickr the coexistence of geographic metadata and text-based annotations (tags) can be exploited to infer new, useful information. This paper introduces a novel method to generate place profiles as vectors of user-provided tags from Flickr geo-referenced photos. These profiles can then be used to measure place similarity in terms of the distance between their matching vectors. A Web-based prototype has been implemented and used to analyze two distinct Flickr datasets, related to a chosen set of top tourism destinations. The system has been evaluated by real users with an online survey. Results show that our method is suitable to define similar destinations. Moreover, according to users, enriching place description with information from user activities provided better similarities.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering

General Terms

Algorithms, Experimentation, Measurement

Keywords

Folksonomies, Geotagging, Place similarity, Flickr

1. INTRODUCTION

Web2.0 and Social Media are becoming crucial in tourism destination online promotion [3]. Within this scenario, pictures shared in social systems are gaining more and more importance [6] because they informally represent and describe destinations. Furthermore, user generated pictures carry a lot of information that could be useful to describe places, as they are often annotated by tags but, at the same time, they also often contain geographic metadata that allow them to be located on a map.

The aim of our work is to exploit image metadata (i.e. tags and geotags) from a social photo sharing system like Flickr to (i) find a representative description for a destination which uses the tags associated to its photos, and (ii) use this description to calculate similarity between destinations and, given a place, suggest new ones which are related to it according to what users say about its buildings, events, and so on. This paper introduces a novel method to generate place profiles as vectors of user-provided tags from Flickr geo-referenced photos. These profiles are then used to measure place similarity. Weights inside vectors are calculated using four different methodologies, that are implemented in a prototype and then evaluated by real users through an online survey.

2. RELATED WORK

Clements et al. [2] introduce a method to predict similar locations (worm-holes) based on human travel behaviour. A wormhole is defined as a similar, but not necessarily spatially close, location on the planet. From a given target location the algorithm aims to find the most similar locations around the world. For each user u, a weight is computed based on the distance of the nearest geotagged photo of the user to the target location. Wormholes are then found by aggregating the geotags of all users with W_u as weight per user, and selecting the most relevant positions on Earth according to this metric. The objectives of this work are clearly the same as ours. However, in our case we measure place similarity on what we call a place *description*, which is built out of tags provided by different users which might, or might not, have visited other places in the world.

Ahern et al. [1] show how to analyze tags associated with geo-referenced Flickr images to generate aggregate knowledge in the form of "representative tags" for arbitrary areas in the world. Data analysis algorithms are based on multi-level clustering and the scoring of tags is based on TF-IDF. With this work we share the idea of tags as representative for a given place, but while the developed tool [1] just shows tags to users, leaving them the interpretation of what they read about a place, our work aims at using that representation to automatically find places which can be considered similar to a given one.

[1] `http://tagmaps.research.yahoo.com/worldexplorer.php`

SAC 2012 March 26-30 2012, Riva del Garda (Trento), Italy.
ACM 978-1-4503-0857-1/12/03.

3. OUR APPROACH

The first step to define similarity between places was to find an appropriate representation for them. We chose to use the Vector Space Model (VSM) [5], representing each place as the weighted collection of all the tags assigned to its pictures.

The folksonomy for a photo sharing system can be defined as a tripartite graph [4] with hyperedges where vertices are partitioned in three (possibly empty) disjoint sets: the set of users $U = \{u_1, u_2, \cdots, u_k\}$; the set of tags $T = \{t_1, t_2, \cdots, t_l\}$; and the set of photos $P = \{p_1, p_2, \cdots, p_m\}$. The folksonomy is thus defined as a set of annotations $A \subseteq U \times T \times P$. Our model also includes a set of destinations $D = \{d_1, d_2, \cdots, d_n\}$ and a geolocation function $coords(\cdot)$ able to return the geographic coordinates of either a destination or a photo. Thus, describing a destination $\hat{d} \in D$ in terms of its associated tags means: (i) building a set of geolocated photos "near" $\hat{d}$: $P_{\hat{d}} = \{\hat{p} \mid \hat{p} \in P, \| coords(\hat{p}) - coords(\hat{d}) \| < \epsilon\}$, where ϵ is a chosen threshold; (ii) building a weighted list of tags from photos in $P_{\hat{d}}$: $T_{\hat{d}} = \{(\hat{t}, w_{\hat{t}}) \mid \hat{t} \in T, \exists \hat{p} \in P_{\hat{d}} : (u, \hat{t}, \hat{p}) \in A\}$, where $w_{\hat{t}}$ is a weight that can be calculated in different ways (e.g. simple or normalized term frequency).

In particular, in our vector space model every destination is represented as a vector of tags $v_{\hat{d}} = \{v_1, v_2, \cdots, v_o\}$, whose size matches the size of the tag vocabulary in the system or of a subset of it (e.g. the most popular tags). Every component in the vector $v_{\hat{d}}$ is either assigned the weight of its matching tag (if it appears in $T_{\hat{d}}$) or zero (if it does not belong to $T_{\hat{d}}$, that is the place has no photos tagged with it).

To test our model we analysed the tags of Flickr geotagged photos of 233 cities, i.e. the union of the top 150 tourism city destinations for the years 2007 and 2008, according to Euromonitor International[2].

We built two different datasets: the former, called *Top100*, contains the 100 most frequent tags for each city, retrieved using Flickr's `flickr.places.tagsForPlace` API. The latter, called *Random*, contains a random sampling of photo metadata (including user- and photo-related information), obtained by querying Flickr APIs with YQL (Yahoo Query Language) and selecting, for each city, 10 photos for 300 random days, taken at random hours so to avoid bias due to day- or time-related events. In both cases, Flickr APIs provide an accuracy parameter which allows to roughly specify the radius around a place's geographic coordinates when selecting photos. In particular, the accuracy ranges from 1 (world level) to 16 (street level). For our datasets, as we deal with cities, we chose to use the city level, which corresponds to the value 11. After a manual analysis of the 1000 most frequent tags in our dataset, we blacklisted all the tags categorized as unuseful (i.e. photography-related terms such as `canon` or `black&white`, or very common terms such as `day` and `dog`) and the ones containing them as substrings. Tags occurring in only one city were also pruned. After cleaning the tag vocabularies, the Top100 and Random datasets counted, respectively, 9,700 and 55,000 distinct tags.

In our model each city is represented by a vector in a multi-dimensional space. Vector components are weighted according to the frequency of the corresponding tags, normalized using the TF-IDF approach. In the Top100 dataset we have calculated standard IDF as $IDF_{\hat{t}} = log_2 \frac{|D|}{|D_{\hat{t}}|}$, where $|D|$ is the number of analyzed places (233), $|D_{\hat{t}}|$ is the number of places containing $\hat{t}$. For the Random dataset two more variants of IDF have been calculated taking advantage of the additional information we were able to retrieve. The former, called *IDFP*, is defined as $IDFP_{\hat{t}} = \frac{|P|}{|P_{\hat{t}}|}$, where $|P|$ is the number of photos in the Random dataset and $|P_{\hat{t}}|$ is the number of photos tagged with $t_{\hat{t}}$. The latter, called *IDFU*, is defined as $IDFU_{\hat{t},\hat{d}} = \frac{|P_{\hat{d}}|}{|P_{\hat{t},\hat{d}}|} \frac{|U_{\hat{t},\hat{d}}|}{|U_{\hat{d}}|}$, where $|P_{\hat{d}}|$ is the number of photos for the destination $\hat{d}$, $|P_{\hat{t},\hat{d}}|$ is the number of photos for the destination $\hat{d}$ tagged with $\hat{t}$, $|U_{\hat{t},\hat{d}}|$ is the number of users tagging destination $\hat{d}$ with the tag $\hat{t}$, and $|U_{\hat{d}}|$ is the number of distinct users who tagged at least a photo in $\hat{d}$.

We calculated tag weights for each of the 233 cities by normalizing tag frequencies in four different ways: (A) Top100 dataset, standard IDF; (B) Random dataset, standard IDF (C) Random dataset, IDFP; (D) Random dataset, IDFU. For each different normalization, we then calculated similarities between cities in terms of the cosine distance between their matching vectors.

4. IMPLEMENTED PROTOTYPE

The prototype we developed to test our model was implemented as a PHP Web-based application, accessing a MySQL database containing the pre-calculated similarity data. The application interface was based on an existing open source project, Geoplanet Explorer[3], and then customized to suit our needs.

Users can select a city from the Top 233 list or specify one manually. In the latter case a disambiguation page is shown, where the user can choose which, among different homonyms, is the place she was referring to (e.g. Rome, Italy vs. Rome in Georgia, US).Once the city has been univocally identified, a Web page is shown containing all the available information about the matching city (see Figure 1): (1) structured data retrieved from Yahoo! GeoPlanet[4]; (2) a small selection of city pictures from Flickr; (3) a map of the city provided by Yahoo! maps; (4) the lists of top 5 similar cities according to the four metrics (System A to D); (5) the most frequent 100 tags for the chosen city. By clicking on one of the top similar cities, a new comparison page opens showing the tags in common between the two. Clicking on one tag at the bottom of the page, instead, brings to anoter Web page showing the top 100 places related to the chosen tag.

5. RESULTS AND EVALUATION

The prototype has been tested with different cities belonging to the dataset and the results are quite interesting. The first thing to note is that, in systems A to C, place similarity is totally independent from who is sharing photos, and it is based only on city descriptions that emerge from the contributions of the whole user community. Tags which represent geographical places, especially in the Random dataset, strongly influence similarity calculation. However this result does not leave us without surprises: the city of Rome, for instance, is considered very similar to Tarragona mainly because of the presence of the tags `Rome`, `Roman`, and so on

[2] `http://tinyurl.com/top150dest`

[3] `http://isithackday.com/geoplanet-explorer/`

[4] `http://developer.yahoo.com/geo/geoplanet/`

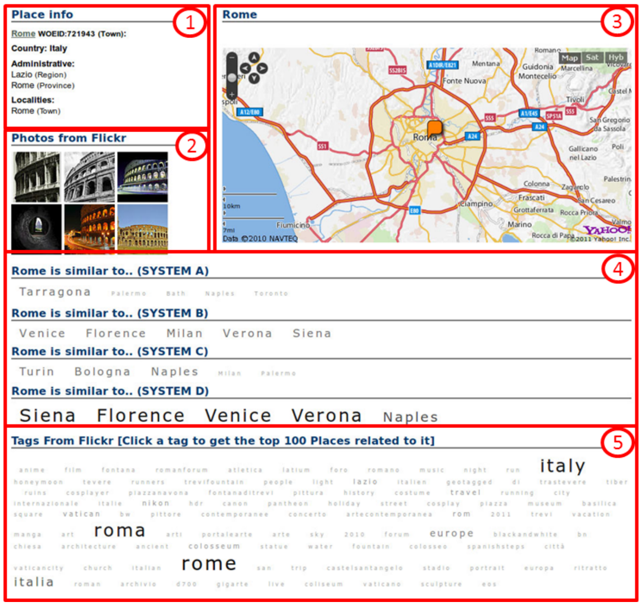

Figure 1: The prototype UI, divided into its five main sections.

(Tarragona has Roman ruins that have been designated a World Heritage Site by UNESCO). Another interesting example is the one of Amsterdam. One of the cities most similar to it is Venice, and the main reason is that canals, boats, and bridges are present in both of them. As a final consideration, it is interesting to see how the main tags characterizing a given location are also the main drivers for the calculation of its similarity with others (e.g. Milan, Turin and Verona are similar thanks to their most characterizing tags `music`, `concert`, and `show`).

To evaluate the system, an online survey was created in order to ask users which of the four similarity measures (System A to D) was the best one according to them. The survey was composed of a demographic section and a group of cities to be evaluated. Users were chosen randomly by posting the survey link on popular social networks. Each user judged one (random) city at a time, for a total of five distinct cities. For each of them, the user was provided results of the four scoring systems, in the form of four lists containing the top-5 related cities (see Figure 1). As the similarity between two places is based on the similarity of their descriptions, users also had the possibility to check the tags in common between two cities.

The survey was completed by 113 users, mostly master students from University of Lugano (Faculty of Communication Sciences) and Politecnico di Milano (Faculty of Computer Engineering), producing 516 valid answers. The best system according to users is System D with 139 preferences. System D is based on IDFU, which gives a higher importance to the user factor (i.e. the higher the number of users agreeing on using a given tag on a given city, the more TF is boosted). System A represents the second most popular answer (118 preferences) and it shows that we can extract valuable knowledge gathering information even just from the Top100 dataset. Conversely, if we compare System A with B –which differs only for the dataset and is ranked fourth– we can see that the choice of the Random dataset alone is not enough to provide better results. The random dataset based on picture information (System C) ranked third according to users' preferences in the survey. Finally System N (with 84 preferences) represents the group with no answers.

Summarizing our results, we can claim that (i) Random dataset has the advantage of providing a richer description for each city, but this description (probably due to the high incidence of geonames and unrelevant words we were not able to catch with our stopwords) is not enough to provide results which are perceived as better; (ii) System A, despite using a less expressive dataset, is able to provide a rather satisfying result with less efforts (i.e. just by using the available Flickr API; and (iii) the fact that System D has been considered the best one allows us to claim that, in our model, user-related information is more relevant for similarity than picture-related information.

6. CONCLUSION

In this paper we have explored the possibility of introducing a new way of measuring place similarity, describing cities in terms of vectors of tags assigned to their photos and calculating their similarities in four different ways. Evaluation of the four systems shows that user-related information is a key factor to improve similarity calculation, while the richness of the tags vocabulary is not as important.

In general, the system has provided interesting suggestions and has been positively evaluated by interviewees. However, we still foresee chances of improvement especially in the following directions: (i) size and variety of the analyzed destinations (using more general "places" instead of top tourism destinations); (ii) controlled enrichment of the tag vocabulary, by automatically filtering geographic tags for individual destinations; and (iii) better selection of the most relevant tags, to allow for more original and serendipitous suggestions.

7. REFERENCES

[1] S. Ahern, M. Naaman, R. Nair, and J. Yang. World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 1–10. ACM, 2007.

[2] M. Clements, P. Serdyukov, A. P. de Vries, and M. J. T. Reinders. Finding wormholes with flickr geotags. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. M. Rüger, and K. van Rijsbergen, editors, *ECIR*, volume 5993 of *Lecture Notes in Computer Science*, pages 658–661. Springer, 2010.

[3] A. Inversini and D. Buhalis. Information convergence in the long tail: The case of tourism destination information. In W. Höpken, U. Gretzel, and R. Law, editors, *ENTER*, pages 381–392. Springer, 2009.

[4] P. Mika. Ontologies are us: A unified model of social networks and semantics. *J. Web Sem.*, 5(1):5–15, 2007.

[5] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[6] K.-H. Yoo, Y. Lee, U. Gretzel, and D. R. Fesenmaier. Trust in travel-related consumer generated media. In W. Höpken, U. Gretzel, and R. Law, editors, *ENTER*, pages 49–59. Springer, 2009.