Università della Svizzera italiana	Facoltà di scienze della comunicazione	lstituto di tecnologie della comunicazione ITC	

Finding Similar Destinations with Flickr geotags



Davide Eynard

University of Lugano ITC - Institute for Communication Technologies Lugano, Switzerland davide.eynard@usi.ch

Alessandro Inversini

University of Lugano ITC - Institute for Communication Technologies Lugano, Switzerland alessandro.inversini@usi.ch

Leonardo Gentile

Politecnico di Milano Milano, Italy leonardo.gentile@gmail.com

Motivations and Objectives

Web2.0 and Social Media are becoming crucial in tourism destination online promotion. Within this scenario, user generated pictures are gaining more and more importance, as:

- they informally represent and describe destinations;
- they bring a lot of related information that could be exploited to better describe places.



Objectives

 build a representative description for a destination, exploiting (geographic and textual) tag information associated to its photos



- This information is mainly provided through:
- geotagging: the process of (explicitly or implicitly) annotating objects and online resources with geospatial context information, typically embedded within picture metadata (stored in EXIF format)
- **tagging**: assigning short strings of text to an individual piece of content in order to classify it for easier retrieval. The collection of terms assigned by all users to all the resources in a system is called *folksonomy*.

Figure 1: a photo of the Lugano lake (ALittaM's Flickr photostream).
Geo coordinates: 46.003661, 8.953342. Tags: lugano, Svizzera,
Switzerland, bianco e nero, wide angle, Sigma 10-20mm,
primavera, città, lago, lake, view, landscape, b&w



 use this description to calculate similarity between destinations, and recommend places which are more similar to a given one

Our approach

1) Cities as bags of tags

given P_d(ε) (set of geolocated photos "near" destination d according to a threshold ε), build a weighted list of tags from photos in P_d
represent every destination as a vector v_d={v₁, v₂,...,v_o}, where o is the tag vocabulary size and every component in v_d is assigned the weighed frequency of its matching tag

2) Data collection and cleaning

233 top tourism destinations according to Euromonitor International (http://tinyurl.com/top150dest) have been considered. After a disambiguation phase, cities are matched to their unique WOEIDs which are then passed to Yahoo APIs for tag retrieval:

3) Weighting

Tag frequencies are normalized by using three different weighting factors: classical IDF (calculated on *Destinations*), IDFP (taking *Pictures* into account), and IDFU (considering both *Pictures* and *Users*)

$$IDF_{\hat{t}} = \log_2 \frac{|D|}{|D_{\hat{t}}|} \qquad IDFP_{\hat{t}} = \frac{|P|}{|P_{\hat{t}}|}$$

$$IDFU_{\hat{t},\hat{d}} = \frac{|P_{\hat{d}}|}{|P_{\hat{t},\hat{d}}|} \frac{|U_{\hat{t},\hat{d}}|}{|U_{\hat{d}}|}$$



4) Recommending

Four different systems:

System name	Dataset	Weighting
Α	Top100	IDF
В	Random	IDF
С	Random	IDFP
D	Random	IDFU

Cosine distance as a similarity measure between destinations:



Amsterdam	Venice	
Amsterdam WOEID:727232 (Town):	Venice WOEID:725746 (Town):	
Country: Netherlands	Country: Italy	
Administrative:	Administrative:	

Veneto (Region)

enice (Province)



Top 1000 frequent tags have been analyzed (see Figure 2) and unrelevant tags have been removed. Two datasets have been generated:

- Top100 (only top 100 tags for each city, vocabulary size 9700 tags)
- Random (tag, user, and photo-related info, 10 photos/day for 300 random days, vocabulary size 55000 tags)



Figure2: distribution of the top 1000 most popular tags, according to our manual classification.



(Town)



Tags in common for Amsterdam and Venice (SYSTEM A)

europa water art night bridge boat canals holiday Canal

Figure3: A comparison between the cities of Venice and Amsterdam, judged similar by System A, and the tags they share.

Our prototype

A prototype has been developed as a PHP Web application, accessing a MySQL database containing pre-calculated similarity data. The interface is based on the opensource project Geoplanet Explorer and is divided in five main sections:



Prototype evaluation

Experimental results

- strong influence of geographic tags (are they really that useful?)
- still interesting surprises (e.g. Rome and Tarragona, Las Vegas and Macau, Milan and Verona/Turin, Venice and Amsterdam – see Figure 3)

Conclusions

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.

Work is still at an early stage and future extensions are envisioned in the following directions: (i) dataset building, (ii) algorithm refinement, (iii) evaluation.

Figure4: Our prototype's main interface, divided in 5 sections.

User evaluation

An online survey was created in order to ask users which similarity measure was the best according to them. Users were randomly chosen by posting the survey link on popular social networks, and they had to judge recommendations provided for 5 cities chosen at random among the 233 top destinations.

During this process, users also had the possibility of checking the tags in common between two related cities.

The survey was completed by 113 users, who produced a total of 516 valid answers. Results are summarized in Fig. 5.

The best system according to users, with 139 preferences, is D, which uses IDFU on the Random dataset. IDFU boosts tags which are used by many users on the same cities.

System A ranks second, showing that we can still extract valuable information from the smaller Top100 dataset (compare with B, which only differs for the dataset).



Scoring Systems

Figure 5: The results of our survey for systems A to D (N=no answer).

References

[1] S. Ahern, M. Naaman, R. Nair, and J. Yang. *World explorer: visualizing aggregate data from unstructured text in geo-referenced collections*. In Proc. 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. Ru ger, and K. van Rijsbergen, editors, ECIR, volume 5993 of LNCS, pagesp 658–661. Springer, 2010.