Mining User Home Location and Gender from Flickr Tags

Adrian Popescu*, Gregory Grefenstette**

*TELECOM Bretagne, France, adrian.popescu@telecom bretagne.eu
**Exalead, France, gregory.grefenstette@exalead.com

Abstract

Personal photos and their associated metadata reveal different aspects of our lives and, when shared online, let others have an idea about us. Automating the extraction of personal information is an arduous task but it contributes to better understanding and serving users. Here we present methods for analyzing textual metadata associated to Flickr photos that unveil users' home location and gender. We test our techniques on a sample of 30,000 people coming from six different countries, allowing us to compare results across cultures and point out similarities and differences.

Introduction and Related Work

With growing acceptance of social computing platforms, some users tend to expose more and more of their lives and their personal data on the Web. Uploading personal photos is a part of the current trend to make personal information available and it is likely to continue developing in the future. Current Flickr user portfolios, consisting of images, textual annotations and photographic metadata span over long periods of time and can reveal a lot about users: their centers of interest, their image of themselves, where they live and what they do. Though some users are careful about explicit disclosure of personal data, they cannot help but reveal some information by the annotations they choose. Examining tagging behavior of 30,000 men and women explicitly declared as residing in six countries, we attempt to discern users' home locations and their gender. One important feature of our work is that we highlight the role of two user characteristics that were neglected in previous studies: location and gender.

There is much new research examining currently developing social computing platforms. (Nov et al, 2009) reported that photo sharing becomes more selective over time and that there is a negative correlation between sharing habits and time. This study addresses the *why* and *how* questions about image annotation but do not treat the questions of *who tags*? and *what is tagged*? The *what* question is addressed by (Sigurbjornsson and van Zwol, 2008), where the authors analyze a large body of Flickr tags to find frequent topics and their correlations. Correlations are then used in order to suggest new tags based on supplied tags.

Our work touches on population specific descriptions of spatial representations, a well studied subject in psychology. (Milgram 1976) showed that landmarks are key components of Parisians' mental representation of their city. Recent research on geotagged photos in photo sharing platforms shows that popular landmarks can be automatically extracted from user annotations (Popescu et al, 2008, Crandall et al, 2009). Among other uses, landmark names can be used to disambiguate tags sets that contain city names. (Lieberman and Lin, 2009) studied Wikipedia contributors who edit geotagged. They mention the identification of the location of Flickr users for future work but seem to conceive an approach that is limited to geotagged content.

(Argamon et al, 2003) provided a detailed analysis of gender identification research over written documents and proposed a method for automatic gender identification based on text characteristics. Their models for "female" and "male" texts were learnt from lexical distributions, allowing new texts to be classified based on these models. (Herring and Paollilo, 2006) have analyzed genre and gender in weblogs and they concluded that language differences are more dependent on the type of blog entry (personal or news commentary) than gender. Since running text is seldom available on photo sharing platforms, these text-based methods are not applicable to image annotations, which are mostly one word tags.

Data Preprocessing and Preliminary Analysis

Since one of our research goals is to compare users located in different countries, we gathered a balanced sample of 5000 users from each of the following countries: *USA, UK, France, Germany, Italy*, and *Spain*. Data were obtained through the Flickr public API, which allows third parties to download information with the user's authorization. An initial problem to solve is the association of users to countries since users disclose their locations in a non-normalized manner. We use lists of synonymous country names in several languages (*United States, Estados Unidos, USA,* etc.) as well as city names (*New York City, Paris, Chicago* etc.) to associate users to locations.

Next, we download photo metadata including photo annotation tags, titles, date taken, upload date, and geotags. Testimonials (description of the target user by other users) also downloaded when available. For each user, a list of unique tags is produced and its elements are then ranked according to their frequency. Preliminary analysis indicates

that, on average, Americans upload the largest number of images in Flickr (1803), followed by British (1398). Italian (585) and Spanish (626) users share the lowest volume of pictures, with around a third of the volume uploaded by Americans. The average size of the tagging vocabulary varies a lot from one country to another (971 for Americans and only 416 for Spanish). (Nov et al, 2009) showed that individual contributions tend to decrease with tenure. The important differences between the contributors from different countries are explained by: the adoption speed (Americans and British were the first to massively create accounts); the commitment to Flickr (224 active days for Americans and only 109 days for Spanish) and the level of detail of the portfolio (Americans seem more willing to provide a more detailed account of their photographic experiences).

Finding a User's Home Location

(Lieberman and Lin, 2009) proposed a method for identifying Wikipedia users' location. We tackle a similar problem for Flickr users and propose a method which exploits manual annotations with place names instead of geotags only. Geotags have the advantage of indicating a image location in an unambiguous way but they are used consistently only by a minority of users. In our sample, only 46.9% of users geotagged photos on at least 3 different days and over a period of at least one month. If only geotags are used, the method is applicable only to a minority of users. Our method is applicable if: Flickr users tag images with the names of the places where photos were taken; place names can be disambiguated; home location is tagged more frequently than other locations and over an extended period of time.

Empirical observation of Flickr tags shows that city names are among the most common location tags used in Flickr and we decided to mine location at a city level. We first build a list of 462 cities from Wikipedia by retaining only cities which have at least 50 associated geotagged articles within a 20 km radius from their point-based coordinates. Each city in the list is described by its name in 5 languages of the users in the sample, the titles of associated geotagged articles (which are often landmarks) and encompassing entities. Some users relocate over a long period of time and in order to minimize the influence of relocations, we search place names only among annotations associated to photos which were taken in the last two years. To remove ambiguities, for a city to be considered, its name (i.e. *Cambridge*) has to be tagged along with one of the following information: one encompassing entity (i.e. *England*) OR one associated landmark (i.e. *King's College*) OR geotags within a 20 km radius (i.e. (52.21N; 0.128E)). To remove candidates which correspond to trips, we retain only cities that were tagged over at least 30 days. For the cities selected according to the above criteria, we count the number of different days (noted *diffDays*) on which they were photographed using the "datetaken" metadata. We assume

that the home location is the one that was photographed on the maximum number of days and, if a tie appears, the city that was photographed over the longest time span is preferred.

		diffDays					
		2	3	4	5	10	20
US	P[%]	81.3	82.5	83.3	84.7	89.6	92.7
	R[%]	75.8	71.1	66.7	63.1	49.1	33.4
UK	P[%]	82.9	83.9	84.9	86.1	90.4	92.9
	R[%]	79.3	74.5	69.7	65.8	49.7	31.8
FR	P[%]	83.9	84.8	85.7	86.4	89.6	92.4
	R[%]	71.4	67.1	62.4	58.4	43.3	26.5
DE	P[%]	80.7	81.7	83.1	83.5	87.5	91.7
	R[%]	63.5	59.1	54.6	51	36.5	22.1
IT	P[%]	85.6	86.7	88.3	89.5	93.6	95.2
	R[%]	77.6	72.9	68	63.4	45.7	25.4
ES	P[%]	80.1	81.4	82.1	83.6	87.4	91.2
	R[%]	65.8	60.8	56.5	52.6	37.4	21.6

Table 1. User location results for: P – correct locations; R – percentage of users for which location was found.

We test our method by comparing automatically discovered locations to locations disclosed on the Flickr profile if the last belong to our list of 462 cities. We present precision and recall results for minimum *diffDays* varying between 2 and 20. The results in table 1 show that for a weak constraint (*diffDays* = 2) location precision is over 80% in all cases, with recall varying from 63.5% (*Germany*) to 79.3% (*UK*). Globally, the best results are obtained for *Italy* and *UK* and the worst for *Germany* and *Spain*. Variations between the six countries in the user sample indicate that British and Italian users are most likely to share photos of their home cities whereas German and Spanish users tag their home locations more rarely. Recall results validate our choice of using tags and geotags instead of geotags only because the maximum recall that could have been obtained with geotags only is under 50%. When increasing *diffDays*, precision improves but recall worsens because location detection is increasingly selective. Results for large values of *diffDays* prove that the main hypothesis underlying our method (i.e. people tend to annotate their home location more often than other locations) is verified. Precision varies for the different countries in the user sample, with best values obtained for Italy (between 85.6% and 95.2% with corresponding recall values 77.6%, respectively 25.4%) and worst results for Spain (between, 80.1% and 91.2% with corresponding recall values of 65.8% and 21.6%).

We have also tried other measures than the count of different days in order to select the home city. Among these, we cite the maximum time span between two photos tagged with a city name or a combination of time span and of different days but results were inferior to those reported above. Automatic user location is possible when people take and tag photos of their home city frequently enough. Errors verification is difficult but errors probably occur for users who: do not photograph their home town; frequently travel to a same location; live between two cities; moved from one city to another without updating their profile.

Finding User's Gender from Tags

We introduce a gender identification technique based on information introduced by other users, "self" photographs and a bag of words inspired method, which are modeled as a cascade classifier. The first classification stage exploits testimonials about the target user which are sometimes present on the user's profile page. These testimonials often include third person pronouns which allow the gender identification. A short list of such pronouns (*she, her,* respectively *he, his* in English; *elle,* respectively *il* in French) is built for each language (adding English pronouns for other languages) and we count the number of female and male pronouns in the testimonials. If one of the counts is bigger than the other, the user is classified accordingly. If no decision is made based on testimonials, we select self-referential photos, i.e. those which are tagged with at least one of the following words: *me, self-portrait, i, myself, selfie* for American and British users; *moi, je, autoportrait* (plus English words) for French. After selecting photos with self-referential tags, we look for tags on the same photo indicating gender such as: *woman, girl* (female); *man, boy* (male) in English; *femme, fille,* respectively *homme, garçon* in French etc. If such gender related words are co-located with self-referential tags, this is a strong indication of the user's gender and we use these correlations to guess one's gender.

Testimonial and self-referential based classifications are very precise but the recall insured by these methods is small (under 20%) and a more generic method is needed to improve recall. Tags reflect what users consider important in their pictures but also reveal the photographer's centers of interest. We hypothesize that male and female tagging vocabularies are different to some extent and that this difference can be used to identify one's gender. To test our hypothesis, we select randomly 1000 users with explicitly declared gender for each country (500 females and 500 males) and build female and male tagging vocabularies. We compute the importance of a tag in a gender vocabulary by counting the number of different users of that gender (out of 500) who used the respective tag and order tags by this frequency of use. Most tags appear in both female and male vocabularies but order differs. For instance, *roses* is ranked 154th for females and 704th for males whereas *panorama* is ranked 1276th for females and 195th for males. We compute tag predominance score

(division of the number of users of one gender who used the tag by the users of the other gender) and present top 10 items for which predominance scores are at least 2:

- *Female*: roses, cookies, jewelry, toes, necklace, cupcakes, valentinesday, yummy, cupcake, haircut
- *Male*: panorama, longexposure, hdr, skyscraper, speed, 50mm, lens, cityscape, jet, canyon

Predominantly female tags are personal (*roses, toes, necklace, valentinesday*) whereas predominantly male tags are more neutral, technical terms (*panorama, hdr, lens*). These results confirm those in (Argamon et al, 2003), where the authors state that texts written by females are more personal than those written by males and we think that they deserve further investigation.

Gender vocabularies are very large and, to speed up processing, we classify gender based on the top 500 elements for both female and male vocabularies. Tag ranks in the two vocabularies are exploited to differentiate between male and female users. Each user's top 500 tags are compared to the two gender vocabularies and we compute a similarity measure using the following relation:

$$sim(U_x, Voc) = \sum_{i=1}^{N} \frac{1}{log_2(rank_x(term_i) * rank_{Voc}(term_i))} \quad (1)$$

Where: U_x - current user's tags; Voc - gender vocabulary; $term_i$ - common term between the two sets of tags; $rank_x$ and $rank_{Voc}$ - ranks of $term_i$ in U_x, respectively Voc.

The similarity score is composed of the common elements between a user's tags and terms in the female and male vocabulary. Basically, if a user's tags are better represented in one gender vocabulary compared to the other, the user is considered to be of the first type. A division by zero is encountered if a term is ranked first in the two sets of tags and, in such cases, the ranks product is set to 2. The contribution of a term to the similarity score decreases with its importance in the two sets of tags via the product of the ranks. To keep individual tags' contributions in a relatively small range, we smooth them using the logarithm of the ranks product. For each user, we obtain a "female" and a "male" score and we can reasonably suppose that the classification performance is higher when the difference between the two scores (*diffGender*) is high. *diffGender* is used a as threshold in order to assess the degree of trust of the method.

The results in table 2 show that when increasing minimum *diffGender*, the classification accuracy increases too but this gain is accompanied by a loss of recall. For *diffGender* = 0.03, precision varies between 85.6% (for Italians) and 89.6% for Americans while recall varies inversely (55.7% for Italians and 48.6% for Americans). The gender difference identification accuracy is important for all values of *diffGender* and all countries (it reaches 76.6% for American female users and 95.2% for American males).

Prior to testing the similarity measure in (1), we tested other similarity measures, such as cosine distance or simple intersection between the two sets of tags but the results were unconvincing.

		diffGender				
		0	0.005	0.01	0.02	0.03
US	F [%]	71.3	72.3	73.6	74.3	76.6
	M[%]	89	90.8	92	94.1	95.2
	All[%]	82.9	84.5	86	87.9	89.6
	R[%]	100	90.1	82.5	68.9	48.6
UK	F [%]	64.9	66.1	66.6	67.9	69.4
	M[%]	83	85.4	87.4	90.1	92
	All[%]	77.9	80	81.7	84.3	86.1
	R[%]	100	93	87.3	76.3	53.4
FR	F [%]	71.6	71.8	72.8	74.9	76.9
	M[%]	79.7	83.2	84.9	87.9	89.7
	All[%]	77.3	79.8	81.3	84.3	86.1
	R[%]	100	85.8	76.9	60.1	47.1
DE	F [%]	65.6	66.8	67.3	68.6	70.1
	M[%]	81.8	84.3	86.4	89.4	91.4
	All[%]	77.1	79.3	81	83.8	85.7
	R[%]	100	84.4	79.4	65.5	53.3
IT	F [%]	66.3	67.2	67.8	69.1	70.3
	M[%]	81.4	84	85.9	89.1	91.3
	All[%]	77	79.2	80.8	83.6	85.6
	R[%]	100	89.4	81.7	66.4	55.7
ES	F [%]	66.9	67.6	68.5	70.2	71.8
	M[%]	80.2	83.4	85.4	88.8	91
	All[%]	76.3	78.8	80.56	83.6	85.7
	R[%]	100	86.3	78.2	62.7	52.6

Table 2. Gender identification results. M – correct identifications for males; F. – correct identifications for females; All – overall correct identifications; R. – recall.

Automatic gender identification is highly effective when users get testimonials or tag their self-referential photos with gender related words but one's tags can also be used to identify her/his gender. Errors are more frequent for females than for males and one could say that males are more predictable taggers than females. We use profile information as ground truth but they can be misleading. For instance, Flickr accounts can be shared by a couple and then the identification of the tagger's gender is very difficult. Also, we build gender vocabularies to account for the average behavior of taggers but this average behavior is misleading in some cases.

Conclusion

We analyze a large sample of Flickr users and focus on their location and gender to mine personal information. This work focuses on two important personal data: location and gender and we show how to extract such information automatically with good accuracy. Future work will focus on improving our methods with the use of more advanced data mining techniques and on discovering users' age.

Acknowledgment

References

Argamon, S., Koppel, M., Fine, J and Shimoni, A. R. 2003. Gender, genre, and writing style in formal written texts. *Text* 23 (3), 321-346.

Crandall, D., Backstrom, L., Huttenlocher, D. and Kleinberg, J. 2009. Mapping the World's Photos. In Proc. of *WWW 2009*, Madrid, Spain, April 2009.

Herring, S. C., and Paolillo, J. C. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics, 10* (4), 439-459.

Lieberman, M. D. and Lin, J. 2009. You are where you edit: Locating Wikipedia contributors through edit histories. *Proc. of ICWSM 2009*, San Jose, CA, May 2009.

Milgram, S. Psychological Maps of Paris. 1976. In *Environmental Psychology: People and Their Physical Settings, 2nd ed*. Holt, Rinehart and Winston, New York, USA, 1976, 104 124.

Nov, O., Naaman, M. and Ye, C. 2009. Motivational, Structural and Tenure Factors that Impact Online Community Photo Sharing. *Proc. of ICWSM 2009*, San Jose, CA, May 2009, 106-113.

Popescu, A., Grefenstette, G. and Moëllic P.-A. 2008. Gazetiki: automatic construction of a geographical gazetteer. *Proc. of JCDL 2008*, Pittsburgh, PA, June 2008.

Sigurbjornsson, B. and van Zwol, R. 2007. Flickr Tag Recommendation based on Collective Knowledge. *Proc. of WWW 2008*, Beijing, China, April 2008.