Mining One Hundred Million Creative Commons Flickr Images Dataset to Flickr Tourist Index

Tao Mao

***Abstract*—Users are continuously uploading and sharing photos these days via many photo-sharing services, for example, Flickr and Instagram. The metadata of the photo data usually contains time-, location- and context-related information. This opens a door for researchers to study human social and/or physical behaviors with different perspectives on different data that they may be interested in. The paper first introduces the newly released "100M Yahoo Flickr Creative Commons Images" dataset for research and briefly describes the information contained in the dataset and its potential applications. The objective of this study is to find most visited place in US by Americans based on geo-tagging information retrieved from the dataset. It proposes a method to detect people's travel patterns as outliers to users' baseline locations. Detected travelling activities are then attributed to the corresponding geo-grids to build Flickr Tourist Index, from which a ranking of most visited places in US is constructed on a yearly basis. Intuitive map visualizations of Flickr Tourist Index are presented on a US map. The paper also studies trends of ranking changes over years and compares its ranking results with other sources.**

***Index Terms*—Big data, photo-sharing service, pig latin, Flickr tourist index.**

I. Introduction

An increasing number of photos are uploaded via social photo-sharing services, e.g., Flickr and Instagram. They often provide datasets or API's to retrieve photos for research purpose. The metadata of these photos usually has geo-tagging information, which comes either from user input or directly from photo capturing devices. This important spatial information can be used to analyze some characteristics of people's activities, such as spatio-temporal tourist dynamics [1]-[3], trip recommendation [4], and hyper-local events [5].

Girardin *et al*. retrieved photographic activity data in the Province of Florence, Italy, via Flickr API, transformed it to spatio-temporal patterns, and derived visualization of user flux [1], [2]. Silva *et al*. compared different abilities of Instagram and Foursquare to provide insights on finding popular places and showing temporal patterns and typical spatial routes [3]. Clement and Serdyukov proposed a kernel convolution method that recommends "wormholes" (similar locations) based on travel patterns obtained from Flickr geo-tagging data [4]. Arase *et al*. argue that uploaded photos indicate people's memorable events since they take efforts to capture and share. They used both geo-tagging information and image tags and titles to detect people's frequent trip patterns and categorize trip types [5].

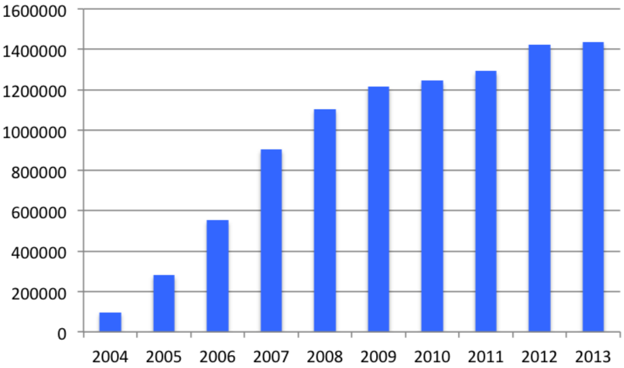


Fig. 1. Number of photos on Flickr captured from 2004 to 2013.

This paper uses "100M Yahoo Flickr Creative Commons Images" dataset for research newly released in June 2014 [6] (below: YFCC100M for short). The dataset of 12GB in size consists of 100 million images compressed in bz2 format, of which 99.3 millions are photos and 0.7 millions are videos. Each image data is expressed as a row record, whose attributes are the following: photo id, user id, captured time, device, title, description, user tags, machine tags, longitude, latitude, geo accuracy, image URL, license, and photo/video marker. Note that geo accuracy 1 represents world level, 3 country level, 6 region level, 11 city level, and 16 street level. We looked at some attributes and obtained some interesting insights: the dataset comes from 581099 users and 27394 device types. Of 48469829 geotagged images, 45857928 images have geo accuracy of city level or more accurate, of which 17170906 were captured in United States. The dataset can also be grouped into subsets by captured year. Flickr was founded in February 2004. Fig. 1 shows the growth of the Flickr adoptions from 2004 to 2013.

The objective of this research is to quantitatively identify from within 48 US contiguous states the top tourist places that Americans visit during holidays from mining the YFCC100M dataset. Thus, we are particularly interested in the following attributes in the data: captured time, longitude, latitude, and accuracy. Attribute "captured time" tells when a user clicks the camera shutter while visiting a place. Also, only images with at least city-level accuracy are considered valid for this research project.

II. Approach

As an outline, we first preprocess the YFCC100M dataset based on time, location and accuracy information, find out

Manuscript received November 4, 2014; revised January 10, 2015.

Tao Mao is with the School of Information at the University of California, Berkeley, USA (e-mail: tao.mao@berkeley.edu).

DOI: 10.7763/IJFCC.2015.V4.365

users' baseline locations, detect their holiday travels as outlines to baseline, and attribute travel activities to geo-grids for building Flickr Tourist Index. (See Appendix A for implementation details)

A. *Data Preprocessing*

With regard to this research purpose, a geographical mask of a rectangular (longitude: W124 - W66; latitude: N24 - N49) is first applied to obtain photo data that represent 48 US contiguous states (excluding Hawaii and Alaska). In addition, the data in recent years from 2010 to 2013 is used for further study since the data in 2014 is not complete. We also filter out data with below city-level accuracy.

B. *Travel Pattern Detection*

We need to determine baseline locations where users live and work, based on the assumption that they would not travel very often during non-holidays. Holidays are defined as U.S. holidays listed on the U.S. Office of Personnel Management website [7]. Holiday blackout dates are the weeks with holidays as well as Saturdays and Sundays before and after these weeks. Thus, any days that are not holiday blackouts are considered non-holidays in this paper. Applying this temporal filter, we can obtain a set of non-holiday locations for each user and calculate the baseline location as the location of medians in longitude and latitude.

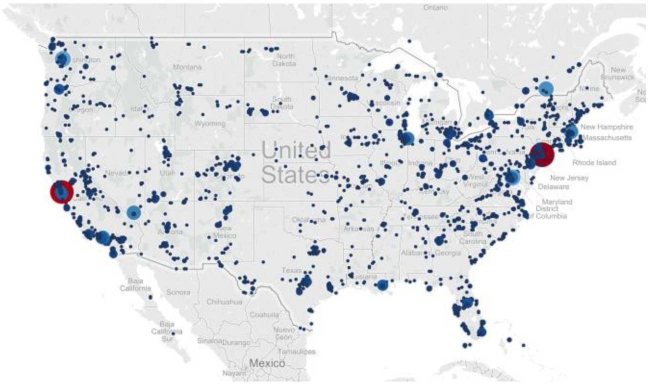
The set of holiday location data (note that it is different from holiday blackouts) is then examined for detecting holiday travels. We simply use Euclidean distance in longitude and latitude and consider 5 degree displacement from baseline (roughly 500 km on the earth surface) as the criterion to determine travel activities.

Travel activities are attributed to the corresponding geo-grids comprised of every 0.1 degree in longitude and latitude. The number of travellers in each grid is defined the Flickr Tourist Index, indicating the tourist attractiveness of a location.

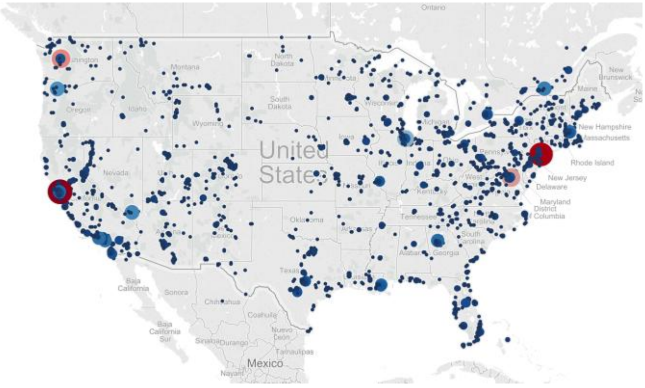
III. Flickr Tourist Index Results

A. *Weighted Map to Show Most Visited Places in US*

We use Pig program to run the computational method described in section II. The weighted map in Fig. 2 shows raw Flickr Tourist Index of places in the US continent in 2012 and 2013. Bubble size and color indicate the intensity of Flickr Tourist Index.



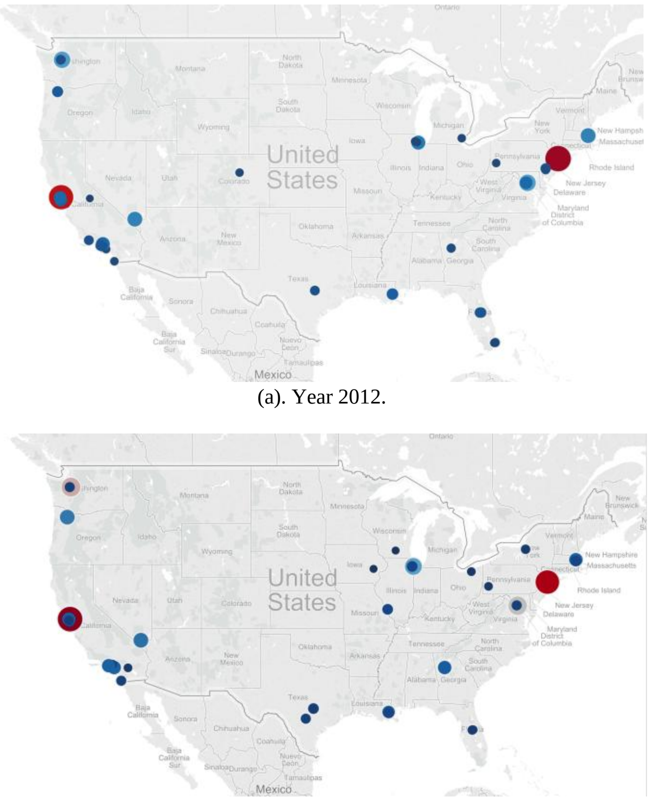
(a). Year 2012.



(b). Year 2013.

Fig. 2. Raw weighted map indicating Flickr Tourist Index of places in US continent in 2012 and 2013 derived from the YFCC100M dataset.

There is one problem on the original map shown in Fig. 2: some locations in Canadian territory are also shown on the north border of United States. In addition, since we are only interested in the yearly most visited places in US, we filter the raw weighted map to get only top ranked places within US. From the filtered map in Fig. 3, we can recognize popular places, e.g., in year 2012, New York City, San Francisco-Bay Area, Las Vegas, Boston, Washington DC, Los Angeles, Chicago, Orlando, and Seattle. and in year 2013, San Francisco-Bay Area, New York City, Los Angeles, Seattle, Washington DC, Chicago and Portland.



(a). Year 2012.

(b). Year 2013.

Fig. 3. Filtered weighted map of the year 2012 and 2013 most visited places in US derived from the YFCC100M dataset.

B. *Are the Rankings Changing over Time?*

Naturally, there is a question to ask: are the rankings of most visited places changing over time? In order to answer this quantitatively, we need to build a more place-friendly mapping of Flickr Tourist Index. City and national park names are identified as their projections on the US map.

Since the current locations are described in longitude and latitude, Flickr Tourist Index should be aggregated to metropolitan areas and major national parks. We decided to use Google map API services to manually do this step because there are a small number of places involved for this step. Finally, we can compile a yearly tourist ranking based on the list of place names and their Flickr Tourist Index scores. Table I summarizes yearly tourist rankings from 2011 to 2013.

TABLE I: YEARLY RANKINGS BASED ON FLICKR TOURIST INDEX FROM 2011 TO 2013

Rank	2011	2012	2013
1	NYC	NYC	SF Bay Area
2	SF Bay Area	SF Bay Area	NYC
3	Boston	Los Angeles	Los Angeles
4	Chicago	Seattle	Seattle
5	Seattle	Washington DC	Washington DC
6	Washington DC	Chicago	Chicago
7	Los Angeles	Boston	Portland
8	Las Vegas	Las Vegas	Boston
9	Portland	Orlando	New Orleans
10	Atlanta	Portland	Las Vegas
11	Austin	Philadelphia	Austin
12	Orlando	Miami	Atlanta
13	Miami	New Orleans	Orlando
14	San Diego	Austin	St. Louis
15	New Orleans	Santa Barbara	San Diego
16	Pittsburgh	Atlanta	San Antonio
17	Nashville	San Diego	Syracus
18	Kansas City	Pittsburgh	Palm Springs
19	Newport (RI)	Denver	Pittsburgh
20	St. Louis	Detroit	Madison

The ranking trends of most visited places from 2010 to 2013 can also shown as chart plots as in Fig. 4: New York City and San Francisco-Bay Area are ranking very stable; Boston and Las Vegas are losing tourists in recent years; Seattle and New Orleans are chasing up while others have ups and downs.

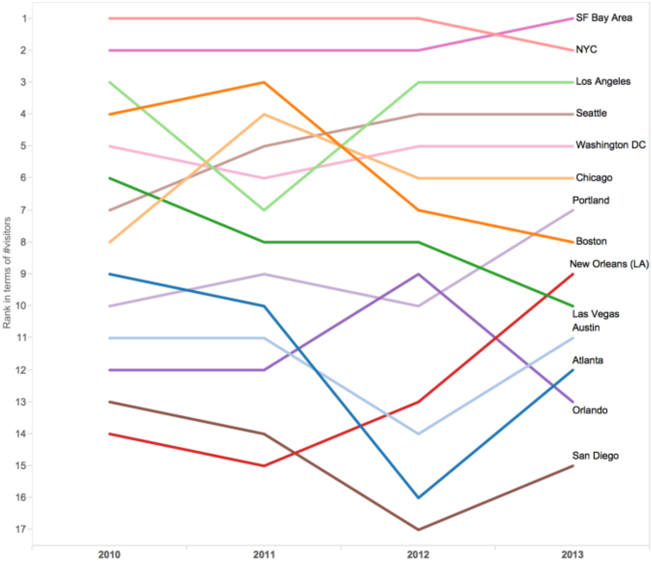


Fig. 4. Ranking trends of most visited places in US from 2010 to 2013.

C. Comparison with Other Ranking

We also compare our 2012 ranking with that from Business Insider [8] in Table II. The two rankings have 8 cities in common out of top 10 and 16 cities out of top 20. The difference can be explained by different methodologies in the two indices: our approach uses photographical activities as an indicator while Business Insider replies on hotel occupancy and may count in business travels.

TABLE II: COMPARISON OF THE YEAR 2012 RANKING OF MOST VISITED PLACES IN US WITH THAT FROM BUSINESS INSIDER

Rank	From this paper	From *Business Insider*
1	NYC	Las Vegas
2	SF Bay Area	NYC
3	Los Angeles	Orlando
4	Seattle	San Diego
5	Washington DC	Los Angeles
6	Chicago	Chicago
7	Boston	SF Bay Area
8	Las Vegas	Washington DC
9	Orlando	Houston
10	Portland	San Antonio
11	Philadelphia	Atlanta
12	Miami	Boston
13	New Orleans	Miami
14	Austin	New Orleans
15	Santa Barbara	Dallas
16	Atlanta	Austin
17	San Diego	Denver
18	Pittsburgh	Philadelphia
19	Denver	Seattle
20	Detroit	Anaheim

IV. CONCLUSION AND FUTURE WORK

This study is believed to be the first work on the YFCC100M dataset to derive tourist insights from spatial information at a large scale. It explores the possibility of mining the whole historical images dataset with geo-tagging information to find popular tourist places in US over the years. The data mining results can serve as tourism guidance to travellers, as well as help cities and states better lay out their recreation and transportation planning by providing insights into tourism hotspots and seasonal trendings.

Since we deal with geographical information, a more sophisticated method - von Mises-Fisher distribution [9] - can be applied to estimate a user's "baseline" especially at a global scale. The Yahoo! Company is planning to provide more visual and audio features of the same dataset, which can potentially be utilized to detect travelling patterns more accurately [6].

APPENDIX

Algorithm Pseudo-Code in Pig Latin

```
-- Load data
-- D - Data
D1 = LOAD '$INPUT' USING PigStorage('\t');

-- Data preprocessing and cleaning
D2 = FOREACH D1 GENERATE *, ISOToDay(ISOTime)
as ISODay;
D3 = FILTER D2 BY ISODay is not null;
D4 = FOREACH D3 GENERATE *, CONCAT(nsid,
CONCAT('@', ISODay)) as nsid_day;
D5 = FILTER D4 BY nsid is not null;

-- Bounded by US continent
D6 = FILTER D5 BY accuracy >= 11 AND latitude >= 24
AND latitude <= 49 AND longitude >= -124 AND longitude
<= -66;
```

```
-- Data split into Nonholiday & Holiday
-- N - Nonholiday
N1 = FILTER D6 BY (UnixTime < $NewYearWeekStart
OR UnixTime >= $NewYearWeekEnd) AND (UnixTime <
$MLKWeekStart OR UnixTime >= $MKLWeekEnd) AND
(UnixTime < $PresidentWeekStart OR UnixTime >=
$PresidentWeekEnd) AND (UnixTime <
$MemorialWeekStart OR UnixTime >=
$MemorialWeekEnd) AND (UnixTime <
$IndependenceWeekStart OR UnixTime >=
$IndependenceWeekEnd) AND (UnixTime <
$LaborWeekStart OR UnixTime >= $LaborWeekEnd) AND
(UnixTime < $VeteransWeekStart OR UnixTime >=
$VeteransWeekEnd) AND (UnixTime <
$ThanksgivingWeekStart OR UnixTime >=
$ThanksgivingWeekEnd) AND (UnixTime <
$ChristmasWeekStart OR UnixTime >=
$ChristmasWeekEnd);

-- H - Holiday
H1 = FILTER D6 BY (UnixTime >= $NewYearDayStart
AND UnixTime < $NewYearDayEnd) OR (UnixTime >=
$MLKDayStart AND UnixTime < $MLKDayEnd) OR
(UnixTime >= $PresidentDayStart AND UnixTime <
$PresidentDayEnd) OR (UnixTime >= $MemorialDayStart
AND UnixTime < $MemorialDayEnd) OR (UnixTime >=
$IndependenceDayStart AND UnixTime <
$IndependenceDayEnd) OR (UnixTime >= $LaborDayStart
AND UnixTime < $LaborDayEnd) OR (UnixTime >=
$VeteransDayStart AND UnixTime < $VeteransDayEnd)
OR (UnixTime >= $ThanksgivingDayStart AND UnixTime
< $ThanksgivingDayEnd) OR (UnixTime >=
$ChristmasDayStart AND UnixTime < $ChristmasDayEnd);

-- Computation
-- NI - Nonholiday_groupby_nsId
NI1 = GROUP N1 BY nsid;
NI2 = FOREACH NI1 GENERATE group, COUNT(N1) AS
c, FLATTEN(Median(N1.longitude)) as mlon,
FLATTEN(Median(N1.latitude)) as mlat;
NI3 = FILTER NI2 BY c >= 20;

-- Median point as baseline
H2 = JOIN H1 BY nsid, NI3 BY group;
H3 = FOREACH H2 GENERATE nsid_day,
ROUND(longitude*10)/10.0 as lon10,
ROUND(latitude*10)/10.0 as lat10, (longitude -
mlon)*(longitude - mlon) + (latitude - mlat)*(latitude - mlat)
AS deviation;

-- Detect travelling pattern and filter out non-travelling data
points
H4 = FILTER H3 BY deviation > 5.0;

-- Sort geo-grid by count
-- HG - Holiday_groupby_Geo
HG1 = GROUP H4 BY (lon10, lat10);
HG2 = FOREACH HG1 {unique_nsid = DISTINCT
H4.nsid_day; GENERATE FLATTEN(group), COUNT(H4)
as c, COUNT(unique_nsid) as uc;};
HG3 = ORDER HG2 BY uc DESC, c DESC;

-- Data storage
STORE HG3 INTO '$OUTPUT' USING PigStorage('\t');
```

Acknowledgment

The following open source packages are used for Pig Latin programming: Apache PiggyBank [10] and Apache DataFu [11].

References

[1] F. Girardin, F. D. Fiore, J. Blat, and C. Ratti, “Understanding of tourist dynamics from explicitly disclosed location information,” in *Proc. the 4th International Symposium on LBS and Telecartography*, Hong Kong, China, 2007.

[2] F. Girardin, J. Blat, F. Calabrese, F. D. Fiore, and C. Ratti, “Digital footprinting: Uncovering tourists with user-generated content,” *Pervasive Computing*, vol. 7, no. 4, pp. 36-43, 2008.

[3] T. H. Silva, P. O. S. V. D. Melo, J. M. Almeida, Salles, and A. A. F. Loureiro, “A comparison of Foursquare and Instagram to the study of city dynamics and urban social behavior,” in *Proc. the 2nd SIGKDD International Workshop on Urban Computing*, 2013.

[4] M. Clements, P. Serdyukov, A. P. D. Vries, and M. J. T. Reinders, “Finding wormholes with Flickr geotags,” in *Proc. the 32nd European Conference on Information Retrieval*, pp. 658–661, 2010.

[5] K. Xie, C. Xia, N. Grinberg, R. Schwartz, and M. Naaman, “Robust detection of hyper-local events from geotagged social media data,” in *Proc. the 13th Workshop on Multimedia Data Mining in KDD*, 2013.

[6] D. A. Shamma. One Hundred Million Creative Commons Flickr Images for Research. [Online]. Available: http://yahoolabs.tumblr.com/post/89783581601/one-hundred-million-creative-commons-flickr-images-for. Retrieved 10/13/2014.

[7] U. S. Office of Personnel Management operating status and schedules. [Online]. Available: http://archive.opm.gov/operating_status_schedules/fedhol/2010.asp. Retrieved 08/10/2014.

[8] Here Are The Cities That Americans Love To Visit In The US. [Online]. Available: http://www.businessinsider.com/most-popular-us-travel-destinations-2013-9. Retrieved 10/13/2014.

[9] N. I. Fisher, *Statistical Analysis of Spherical Data*, Cambridge University Press, 1993.

[10] Piggy Bank Apache Official Website. [Online]. Available: https://cwiki.apache.org/confluence/display/PIG/PiggyBank. Retrieved 08/10/2014.

[11] Apache DataFu. [Online]. Available: http://data.linkedin.com/opensource/datafu. Retrieved 08/10/2014.

Tao Mao is a master student of information and data science in the School of Information at the University of California, Berkeley and concurrently a software engineer at Yahoo! Inc. He holds a Ph.D. degree in engineering sciences from Thayer School of Engineering, Dartmouth College in United States and a bachelor degree in electrical engineering from Zhejiang University in China. His research interests include big data, machine learning, reinforcement robotics, and artificial neural networks.