Evaluating Temporal Information for Social Image Annotation and Retrieval

Tiberio Uricchio, Lamberto Ballan, Marco Bertini, and Alberto Del Bimbo

Media Integration and Communication Center (MICC)
Università degli Studi di Firenze, Italy

Abstract. *Can we use the temporal evolution of annotations in Web images to improve tasks such as annotation, indexing and retrieval?* This important question is the main motivation for this work. Typically visual content, text and metadata, are used to improve these tasks. A characteristic that has received less attention, so far, is the temporal aspect of social media production and tagging. The main contribution of this paper is a thorough analysis of the temporal aspects of two popular datasets commonly used for tasks such as tag ranking, tag suggestion and tag refinement, namely NUS-WIDE and MIR-Flickr-1M. The correlation of the time series of the tags with Google searches shows that for certain concepts web information sources may be beneficial to annotate social media.

Keywords: Temporal information, image annotation, image retrieval, image tagging, social media.

1 Introduction

The huge success of sites and applications for creation, sharing and tagging of user-generated media - such as Flickr, Facebook and YouTube - has lead to a strong interest by the multimedia and computer vision communities in researching methods and techniques for annotating and searching social media. Typically visual content, text and metadata, such as geo-tags, are used to improve tasks such as annotation, indexing and retrieval of the huge quantities of media produced every day by the users of such systems. For instance, visual content similarity is used in [15] to perform tag suggestion and image retrieval, tag co-occurrence has been proposed in [19] for tag suggestion, geo-tags have been used in [20] for tag recommendation, content classification and clustering. A recent review of the state-of-the-art in areas related to web-based social communities and social media has been presented in [21], considering in particular the contribution of contextual and social aspects of media semantics to multimedia applications.

A characteristic that has received less attention, so far, is the temporal aspect of social media production. As noted in [2], extracting time information from documents may improve several applications such as hit-list clustering and exploratory search. More recently, several researchers have shown that the temporal information associated to search engine queries (e.g. frequency of query

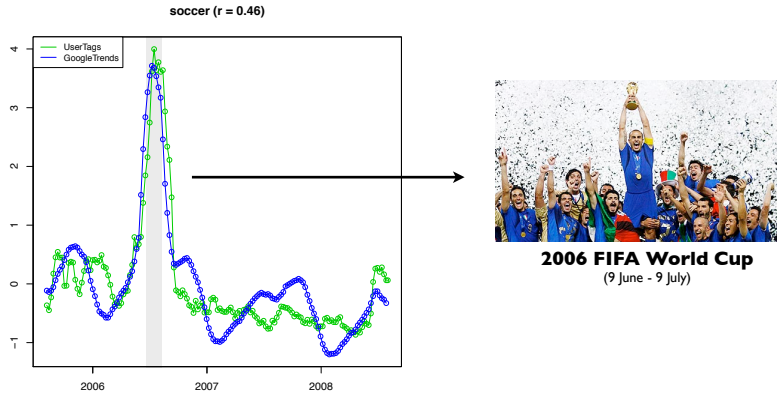

Fig. 1. Time series of user tags and Google searches for "soccer" in NUS-WIDE dataset.

keywords over time) can be used to predict trends and behaviors related to economics and medicine, such as claims for unemployment benefits [4], and detection of flu epidemics [7].

In [18] "burst" analysis techniques derived from signal processing are compared against a novel method to identify social events in the associated social media, using the tags and geo-localization information of Flickr images. In [14], the temporal evolution of topics in social image collections is proposed to perform subtopic outbreak detection and to classify noisy social images. The authors used a non-parametric approach in which images are represented using a similarity network, created using Sequential Monte Carlo, where images are the vertices and the edges connect the temporally related an visually similar images. Temporal dynamics of social image collections has been studied in [13] to improve search relevance at query time, addressing both a general case and personalized interest searches. The authors propose a unified statistical model based on regularized multi-task regression on multivariate point process, in which an image stream is considered an instance of a process and a regression problem is formulated to learn the relations between image occurrence probabilities and temporal factors that influence them (e.g. seasons).

Analysis of the temporal evolution of social media collections have been proposed in [10] to predict political success and product sales; regression-based and diffusion-based models have been adapted to account for a Flickr-based index, combining images' metadata and visual similarity, that models the popularity of politicians and products. The work presented in [12] re-casts the problem of image retrieval re-ranking as a prediction of which images will be more likely to appear on the web at a future time point. Both collective group level and individual user level cases are considered, using a multivariate point process to model a stream of input images, and using a stochastic parametric model to solve the relations between the occurrences of the images and factors such as visual clusters, user descriptors and month of the image.

All the datasets used in these works are based on custom selections of user-generated images selected from Flickr, and are not publicly available. The main

contribution of this paper is a thorough analysis of the temporal aspects of two "standard" datasets commonly used for tasks such as tag ranking, tag suggestion and tag refinement [16] [15] [23] [17] [3]: NUS-WIDE [5] and MIR-Flickr-1M [9]. These datasets provide images and associated metadata, along with a ground-truth annotation of 81 and 18 tags, respectively. Analysis of the temporal evolution of both user tags and ground-truth tags allows to evaluate the social context (e.g. use of tags related to the semantics associated to social interaction, and not necessarily associated with image content) and visual content (e.g. use of tags that are more strictly related to image content). The correlation of the time series of the tags with Google searches (see Fig. 1) shows that for certain concepts web information sources may be beneficial to annotate social media.

2 Data Analysis Method

2.1 Datasets

To measure the impact of temporal information for image annotation purposes, we performed a quantitative analysis over two image datasets: NUS-WIDE [5] and MIR-Flickr-1M [9].

NUS-WIDE is a large scale dataset collected from Flickr. It contains 269,648 images, provided as multiple visual features and source URLs, with 5,018 tags of which 81 have been manually checked and can be considered ground-truth tags. Tab. 2.1 reports the classification of these tags according to their main WordNet category. In order to obtain all temporal metadata not contained in the set, we had to download again all the original images from Flickr. Unfortunately, some images are not available anymore, therefore we had to use a subset of 238,251 images that are still present on Flickr. We refer to this subset as NUS-WIDE-240K. Images are unbalanced with respect to time, having very different number of images per date. The time interval goes from year 1900 (old photo scans) to 2009, concentrating most of the images between 2005-2008.

MIR-Flickr-1M is also a large dataset crawled from Flickr which contains 1 million images, selected by their Flickr interestingness score [1] [8]. Every image provided has full *Flickr metadata* which includes *taken* and *posted* timestamps, indicating when a photo was taken and when it was shared on Flickr. However, only about half of the images provide a valid "taken" timestamp, in particular only 584,892 are valid, as 330,454 have no timestamps and 84,654 have an invalid timestamp. Like NUS-WIDE-240K, images are unbalanced with respect to time. Images are concentrated around years 2007-2009. A ground-truth comprised of 18 tags is provided for the first 25,000 images only, that compose a subset called MIR-Flickr25K [8].

2.2 Temporal features

Given a set of images I, all taken in a set of dates D (as a daily interval), we denote as T the set of all tags used and U the set of all users. For every image

Object	12	Animal	13	Location	2	Substance	2
Action	5	Plant	4	Top	4	Time	2
Artifact	26	Event	4	Phenomenon	4	Person + Groups	3

Table 1. WordNet categories of NUS-WIDE ground-truth tags.

$i \in I$ we denote $\text{tag}(i) \subseteq T$ the set of tags associated, $day(i) \in D$ the timestamp associated and $user(i) \in U$ the user who owns the image. We also consider two other time spans, a set of weeks W and a set of months M, easily computed by integrating over the interval of days considered. These can be thought as time series over the selected index set. For every set considered, we computed a set of features, as proposed in [12]:

- **Images per day**: the number of relevant images which are *taken* in a day. More specifically, given a day $d \in D$, the number of images per day (IMD) is defined as
$$\text{IMD}(d) := |\{i \in I | day(i) = d\}| \tag{1}$$
Similarly we also define a feature for the number of images per week (IMW) and per month (IMM).
- **Images per day for a tag**: the number of relevant images associated with a tag which are *taken* in a day. More specifically, given a tag $t \in T$ and a day $d \in D$, the number of images with t per day (ITD) is defined as
$$\text{ITD}(t, d) := |\{i \in I | day(i) = d \wedge t \in tag(i)\}| \tag{2}$$
Similarly we also define a feature per week (ITW) and per month (ITM).

However, a phenomenon associated with a social source is that of *batch tagging*: a user may decide to upload an entire album of photos and, instead of carefully tagging each photo, he could simply opt to tag each photo with the same tags (e.g. tag the album instead of every single photo). This may result in a kind of noise with respect to the normal use of tags in time. In addition, the features defined above are sensitive to this kind of noise, producing noisy peaks over single days. To produce a more meaningful analysis we decide to collapse all images that are batch tagged into a single entry. A set of images are considered *batch tagged* if they are all uploaded by the same user on the same day and have the same set of tags. More specifically, given a user $\hat{u} \in U$, a day $\hat{d} \in D$ and a set of tags $\hat{t} \subseteq T$, a set of images $I_B = \{i_1, i_2, \ldots, i_k\}$ are considered *batch tagged* if $\text{tag}(i) = \hat{t}, \text{user}(i) = \hat{u}, \text{day}(i) = \hat{d}\ \forall i \in I_B$.

2.3 Flickr Popularity Model

As described in [10], available images from the two datasets are only a sample of all images in Flickr. In addition, the number of images over time in Flickr are mostly variable, based on the popularity of the site itself. This slow change over time can be modeled as a trend over all tags, independent from any particular query. Unfortunately, no statistics are released publicly and other sources such as Alexa[1] or Google Trends[2] are affected by the impact of news. Based on this preliminary analysis and supposing an uniform sampling in Flickr searches, we use the feature IMD to remove this background deviation by normalizing the ITD feature.

[1] Alexa Internet, Inc. `http://www.alexa.com`
[2] Google Trends. `http://www.google.com/trends`

Given a tag $t \in T$ and a date $d \in D$ we compute:

$$\overline{ITD}(t,d) = \frac{ITD(t,d)}{IMD(d)} \tag{3}$$

This may also be considered as a frequentist probability distribution of tag t in day d with respect to all other tags considered, which is $p(t;d)$. Similarly we also compute $\overline{ITW}$ and $\overline{ITM}$ by considering a week and a month granularity, respectively. After collapsing all batch tagged images, the two datasets retain 179,128 images for NUS-WIDE-240K and 531,670 images for MIRFLICKR-1M respectively.

2.4 Processing

First of all we present a qualitative analysis by measuring the occurrence of tags in time. Given that NUS-WIDE-240K has the biggest ground truth of all datasets considered and that we are looking to discover the relations between tags and image content with respect to time, we choose to use it as the main reference. We use all the 81 manually checked tags as T set and consider four different information sources which are different in the kind of underlining latent process :

- From NUS-WIDE-240K, for all images, we consider the T set of tags using the **manually validated** tags which constitute the entire ground truth; we refer to this source as **NUS-GT**.
- From NUS-WIDE-240K, for all images, we consider the T set of tags using the **user tags** (e.g. the tags provided by the respective Flickr users); we refer to this source as **NUS-TAGS**.
- From MIRFLICKR-1M, for all images, we consider the T set of tags using the **user tags**; we refer to this source as **MIR-TAGS**.
- Beside image datasets, we also consider a source of temporal query information given by Google Trends. From Google Trends, we have downloaded all available query data for the T set of tags considered; we refer to this source as **GOO-TAGS**.

All sources are to be considered subject to different kinds of noise, in particular all images are highly unbalanced over time, resulting in days with hundreds of images and others with at most ten images. To reduce this effect, we choose to consider only the largest time span with at least 350 images per week. In addition the two image datasets differ in the time interval which has the most images. This forced us to use a reduced time interval that we choose as starting from 2005-06-01 and ending in 2008-08-01 for NUS-WIDE-240K (retaining 161,176 images from 179,128) and from 2007-01-01 to 2008-08-01 for MIR-Flickr-1M (retaining 110,064 images from 531,670). Those filters were processed with a combination of Python scripts and Google Refine[3]. After this we used the R package [22] to plot and execute any successive analysis. A plotting of features of this data revealed an insufficient reduction in noise to be able to clearly visualize

[3] Google Refine. `http://code.google.com/p/google-refine`

most characteristics pattern. To make the time series patterns more clear, we computed a simple moving average over all time series, varying the windows size n from 2 to 10 weeks. For a day time series defined over a time span Ψ for a tag $t \in T$ is defined as:

$$ITD_n(t,d) = \frac{1}{n}\sum_{i=-n}^{n} \overline{ITD}(t,d+i) \quad \forall d \in \Psi \tag{4}$$

This has the effect to smooth the series, letting to visualize more clearly the trend. On the other hand, tags which have very sparse frequency tends to be worsened, so we adjusted the window size empirically, based on visualization clearness. The final time series are composed of 1,158 and 579 week samples respectively for NUS-WIDE-240K and MIR-Flickr-1M.

2.5 Correlation analysis

To exploit the underlining time process and to be able to improve image annotation using temporal information, we need a way to evaluate quantitatively the possible correlation between sources. This allows us to analyze if a series can be estimated by another one and how a generalized model may describe the original time series. To this end we compute a correlation measure over two series. First of all we standardize all time series: given a time series $X = \{x_i : i \in D\}$, we compute $x_i = \frac{x_i - \overline{X}}{s}$, where $\overline{X}$ is the sample mean and s is the sample standard deviation. Even if sample mean and sample standard deviation are sensible to outliers, those are removed thanks to the filtering and smoothing procedure described above. To evaluate the correlation between two time series, we choose to use the *sample Pearson correlation coefficient*, often denoted as r. Given two time series X and Y of n samples, r is defined as the ratio between covariance and the product of X variance and Y variance:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{Y})^2}} \tag{5}$$

which is defined in $[-1, 1]$. Values towards the positive or negative end reveal a strong correlation between the two time series, changing only in the sign. We can reformulate it as the mean of the products of the standard scores, which permits us to use standardized time series $\hat{x_i} = \frac{x_i - X}{s_X}$ and $\hat{y_i} = \frac{y_i - Y}{s_Y}$:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\Big(\frac{x_i - X}{s_X}\Big)\Big(\frac{y_i - Y}{s_Y}\Big) = \frac{1}{n-1}\sum_{i=1}^{n}\hat{x_i}\hat{y_i} \tag{6}$$

Given that the strength of correlation is not dependent on the direction or the sign, we also computed r-square. Unfortunately the interpretation of a correlation coefficient depends heavily on the context and purposes that can't be easily defined at this stage of work. However several works like [6] offered some guidelines which can be used to interpret our analysis, that are reported in Tab. 2.

Correlation	None	Small	Medium	Strong
Positive	0.0 to 0.09	0.1 to 0.3	0.3 to 0.5	0.5 to 1.0
Negative	-0.09 to 0.0	-0.3 to -0.1	-0.5 to -0.3	-1.0 to -0.5

Table 2. Guidelines for sample Pearson correlation coefficient.

3 Experiments and Discussion

In the following we will consider both the presence of the tags that have been added by the users that uploaded the images to Flickr (referring to them as "user tags") and the tags that have been manually checked by the creators of NUS-WIDE as referring to visual content of images (referring to them as "ground-truth" tags). In fact, several studies have shown that tags are often ambiguous and personalized [11] [19], and do not necessarily reflect the visual content of the image. As an example consider Fig. 2, showing the temporal usage of the tags "snow" and "soccer" in NUS-WIDE, along with the respective Google searches, as obtained from Google Trends. It can be observed that the peak in usage of the "soccer" tag - associated with the 2006 FIFA World Cup - reflects that in Google Trends, but the peak is much less pronounced in the ground truth tags; this indicates that for this tag the relationship between tags and image may exist because of how people react to social events, rather than uploading photos depicting that event on Flickr. On the other hand the peaks of both user and ground truth "snow" tag are corresponding to that of Google Trends: in this case the relationship may exist because it is more likely that people take pictures of snow scenes during winter, and this concept is less related to social aspects than to visual content of these images.

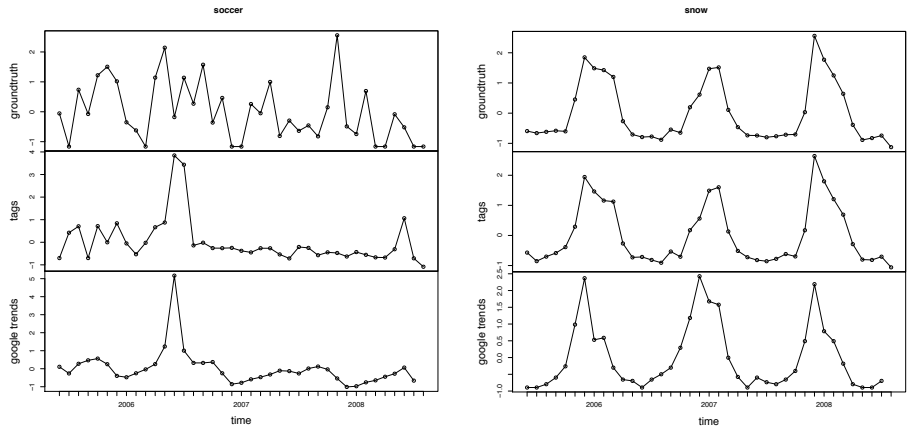

Fig. 2. *left)* frequency of "soccer" in NUS-GT, NUS-TAGS and GOO-TAGS: the peak of Google Trends and user tags in the summer of 2006 are related to the World Soccer Championship; *right)* frequency of "snow" in NUS-GT, NUS-TAGS and GOO-TAGS: the peaks are associated with winter seasons. Tag frequencies have been normalized by the number of images of the same day.

3.1 Temporal Evaluation

Considering time series composed of the frequencies of image tags (either user or ground-truth) and Google searches obtained from Google Trends, it is possible to

observe that they exhibit the presence of different components, that may appear mixed together:

trend long term variation, that can be increasing, decreasing or also stable (see Fig. 3 left). Terms such as "computer" or "military" have this pattern;

cyclical variation repeated but not periodic variations. Tags like "sports" or "flags" have this pattern;

seasonal variation periodic variations, e.g. due to concepts associated with some regular event (see Fig. 3 center). Concepts related to seasons show this behavior, like "garden", "snow", "beach" or "frost";

irregular variation random irregular variations, e.g. due to the sudden emergence of a topic (see Fig. 3 right), that appears as a burst of activity. Concepts that exhibit this pattern are related to social or natural events like "soccer", "earthquake" and "protest".

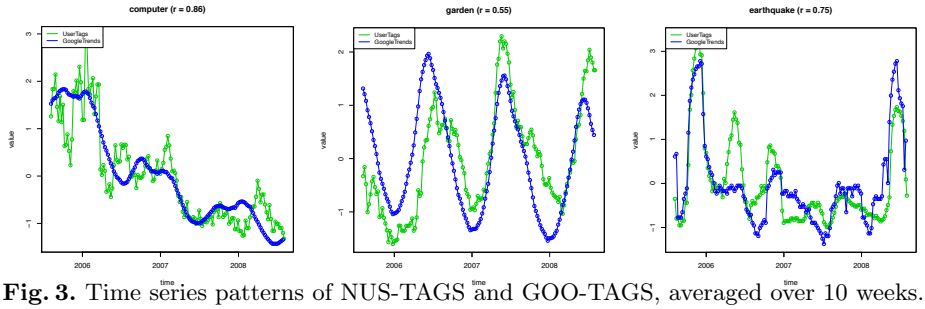

Fig. 3. Time series patterns of NUS-TAGS and GOO-TAGS, averaged over 10 weeks. *left)* trend (computer); *center)* seasonal (garden); *right)* episodic (earthquake: peaks correspond to earthquakes in China and Pakistan).

3.2 Correlation Analysis

Fig. 4 reports the outcome of correlation analysis of NUS-TAGS with NUS-GT, NUS-TAGS with GOO-TAGS and NUS-GT with MIR-TAGS. In particular it can be observed that the correlation of NUS-TAGS and NUS-GT has a vast majority of "Medium" and "Strong" values, while the correlation between user tags and Google searches is overall weaker and can be useful for a selected number of tags. The correlation between NUS-GT and MIR-TAGS has a large number of "Medium" and "Strong" values, suggesting that the temporal information of NUS-WIDE can be used in MIR-Flickr-1M.

Correlation analysis of NUS-TAGS with GOO-TAGS, followed by averaging of r-square values over tags classes (Fig. 5 left) shows that Plant, Event, Phenomenon and Action obtain the higher values. A second group of categories comprises Artifact, Person+Group, Animal, Object and Time. In general, the categories that obtain the best performances are benefitting from tags whose time series show seasonal behaviors (e.g. "snow", "frost", "grass", "leaf") or have a "burst" behavior associated with specific social events (e.g. "soccer", "protest", "earthquake").

Correlation analysis of NUS-GT with GOO-TAGS (Fig. 5 right) shows that Plant and Phenomenon categories maintain their position among the best performing classes, because of the tags that exhibit a seasonal pattern. Instead the

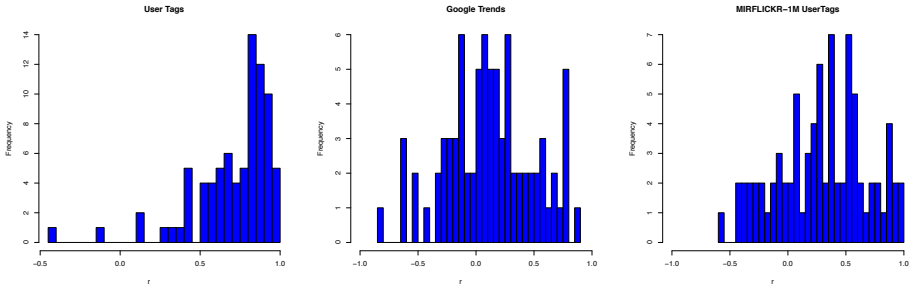

Fig. 4. *left)* r values computed between NUS-TAGS and NUS-GT; *center)* r values computed between NUS-TAGS and GOO-TAGS; *right)* r values computed between NUS-GT and MIR-TAGS.

correlation of Event and Action categories is lower because the ground-truth tags that have an episodic pattern like "soccer", "protest" and "earthquake" have a lower correlation. This is due to the fact that these tags are employed by users also when the content of the image is not visually related to the described event.

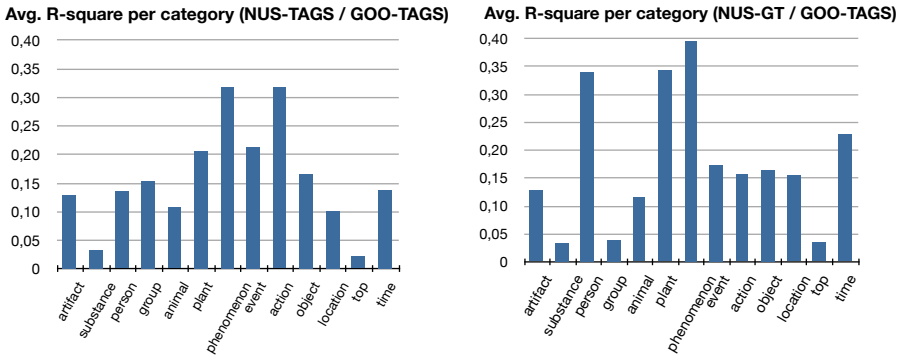

Fig. 5. NUS-WIDE dataset: r-square averages for tags classes. *left)* NUS-TAGS correlation with GOO-TAGS; *right)* NUS-GT correlation with GOO-TAGS.

4 Conclusion

This paper presented a thorough analysis of the temporal aspects of user annotations in two popular large-scale datasets. The correlation of the time series of the tags with Google searches showed that for certain concepts web information sources may be beneficial to annotate social media.

References

1. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proc. of ACM CHI (2004)
2. Alonso, O., Gertz, M., Baeza-Yates, R.: On the value of temporal information in information retrieval. SIGIR Forum 41(2), 35–41 (Dec 2007)

3. Uricchio, T., Ballan, L., Bertini, M., Del Bimbo, A.: An evaluation of nearest-neighbor methods for tag refinement. In: Proc. of IEEE ICME (2013)
4. Choi, H., Varian, H.: Predicting the present with Google Trends. Tech. rep., Google (2011)
5. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: A real-world web image database from National University of Singapore. In: Proc. of ACM CIVR (2009)
6. Cohen, J.: Statistical power analysis for the behavioral sciences. Routledge Academic (1988)
7. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature 457(7232), 1012–1014 (02 2009)
8. Huiskes, M.J., Lew, M.S.: The MIR Flickr retrieval evaluation. In: Proc. of ACM MIR (2008)
9. Huiskes, M.J., Thomee, B., Lew, M.S.: New trends and ideas in visual concept detection: the MIR Flickr retrieval evaluation initiative. In: Proc. of ACM MIR. pp. 527–536 (2010)
10. Jin, X., Gallagher, A., Cao, L., Luo, J., Han, J.: The wisdom of social multimedia: using Flickr for prediction and forecast. In: Proc. of ACM MM. pp. 1235–1244 (2010)
11. Kennedy, L.S., Chang, S.F., Kozintsev, I.V.: To search or to label? Predicting the performance of search-based automatic image classifiers. In: Proc. of ACM MIR (2006)
12. Kim, G., Fei-Fei, L., Xing, E.P.: Web image prediction using multivariate point processes. In: Proc. of ACM SIGKDD. pp. 1068–1076 (2012)
13. Kim, G., Xing, E.P.: Time-sensitive web image ranking and retrieval via dynamic multi-task regression. In: Proc. of ACM WSDM. pp. 163–172 (2013)
14. Kim, G., Xing, E.P., Torralba, A.: Modeling and analysis of dynamic behaviors of web image collections. In: Proc. of ECCV. pp. 85–98 (2010)
15. Li, X., Snoek, C.G.M., Worring, M.: Learning social tag relevance by neighbor voting. IEEE Transactions on Multimedia 11(7), 1310–1322 (2009)
16. Liu, D., Hua, X.S., Yang, L., Wang, M., Zhang, H.J.: Tag ranking. In: Proc. of WWW (2009)
17. Liu, D., Yan, S., Hua, X.S., Zhang, H.J.: Image retagging using collaborative tag propagation. IEEE Transactions on Multimedia 13(4), 702–712 (2011)
18. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from flickr tags. In: Proc. of ACM SIGIR. pp. 103–110 (2007)
19. Sigurbjörnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: Proc. of WWW. pp. 327–336 (2008)
20. Sizov, S.: Geofolk: latent spatial semantics in web 2.0 social media. In: Proc. of ACM WSDM. pp. 281–290 (2010)
21. Sundaram, H., Xie, L., De Choudhury, M., Lin, Y.R., Natsev, A.: Multimedia semantics: Interactions between content and community. Proceedings of the IEEE 100(9), 2737–2758 (2012)
22. Team, R.C.: R: A language and environment for statistical computing. vienna, austria: R foundation for statistical computing; 2008 (2011)
23. Zhu, G., Yan, S., Ma, Y.: Image tag refinement towards low-rank, content-tag prior and error sparsity. In: Proc. of ACM Multimedia (2010)