2012 IEEE 12th International Conference on Data Mining

Towards Automatic Image Understanding and Mining via Social Curation

Katsuhiko Ishiguro, Akisato Kimura, and Koh Takeuchi
NTT Communication Science Laboratories
NTT Corporation, Kyoto, Japan
ishiguro.katsuhiko@lab.ntt.co.jp, akisato@ieee.org

***Abstract*—The amount and variety of multimedia data such as images, movies and music available on over social networks are increasing rapidly. However, the ability to analyze and exploit these unorganized multimedia data remains inadequate, even with state-of-the-art media processing techniques. Our finding in this paper is that the emerging *social curation* service is a promising information source for the automatic understanding and mining of images distributed and exchanged via social media. One remarkable virtue of social curation service datasets is that they are weakly supervised: the content in the service is *manually collected, selected and maintained* by users. This is very different from other social information sources, and we can utilize this characteristics for media content mining without expensive media processing techniques. In this paper we present a machine learning system for predicting view counts of images in social curation data as the first step to automatic image content evaluation. Our experiments confirm that the simple features extracted from a social curation corpus are much superior in terms of count prediction than the gold-standard image features of computer vision research.**

***Keywords*-Social curation, automatic image understanding and evaluation, feature extraction, regression**

I. Introduction

The amount and value of information distributed and consumed in social networking services (SNS) are increasing rapidly in both business and academic settings. At the same time, there are urgent demands to develop an efficient and intuitive method to identify the interesting and valuable information from the growing volume of social media content such as texts , images, movies and music [1], [2], [3], [4].

In this setting, *social curation* service is emerging as a new way to interact with social media. At the most basic level, curation service offers the ability to (i) bundle a collection of content from diverse sources, (ii) re-organize them to give one own perspective, and (iii) publish the resulting story to consumers. In Fig. 1, we present a schematic view of curation service. A curator organizes *a curation list*, which is a compilation of social media content, from a pool of content created by others. It is worth noting that a curation list is a collection of social media content that is *manually collected, selected and maintained* by its curator.

As described later, social curation service and curation lists contain a lot of images and movies. The former is particularly attractive to many SNS users. Therefore, an automatic scheme for understanding and evaluating image content on SNSs would be beneficial for improving click-through rates of the social curation service and the effectiveness of advertisements. However, the automatic understanding of images, one of the ultimate goals of computer vision research, remains quite difficult for the unorganized image content in the Web and SNSs, even for state-of-the-art techniques (e.g. [5], [6], [7]).

The main claim of this paper is that we are able to understand and mine images in SNSs by utilizing social curation data. We assume that the contents of a curated list, including image content, are manually organized to fully convey the curators intentions. It follows that we can infer the context or evaluation of an image from the social information and contextual features of the other (non-image) content in the curation list, and the curation list itself; the goal is to dispense with expensive computer vision techniques. The only technical study related to social curation service is related to text corpus analysis [8]. To the best of our knowledge, this is the first work to examine social curation data for automatic image understanding and mining.

The rest of the paper is organized as follows. In the second section, we review related works. In the third section, we explain the social curation service, our target data source, and detail the dataset specifications. The fourth section is devoted to the problem formulation of predicting view counts of an image included in a curation list, as a first step towards image understanding and mining via social curation. The fifth section describes experiments conducted to confirm the effectiveness of the social curation information compared to the current gold-standard features used in computer vision research. The last section concludes this paper with a discussion about future works.

II. Related Works

One notable trend in SNS-related research is agglomerating multiple information sources or services to obtain a deeper understanding of social media content. For example, Mejova and Srinivasan [9] employ a domain adaptation technique for sentiment analysis in three different social media streams: weblogs, review articles, and tweets on Twitter. The authors of [10] extend a topic model [11] to associate tweets and real events to discover topical segmentation in a event. Kulshrestha et al. [12] studied the impact of offline geolocations on online social network activities and

1550-4786/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDM.2012.37

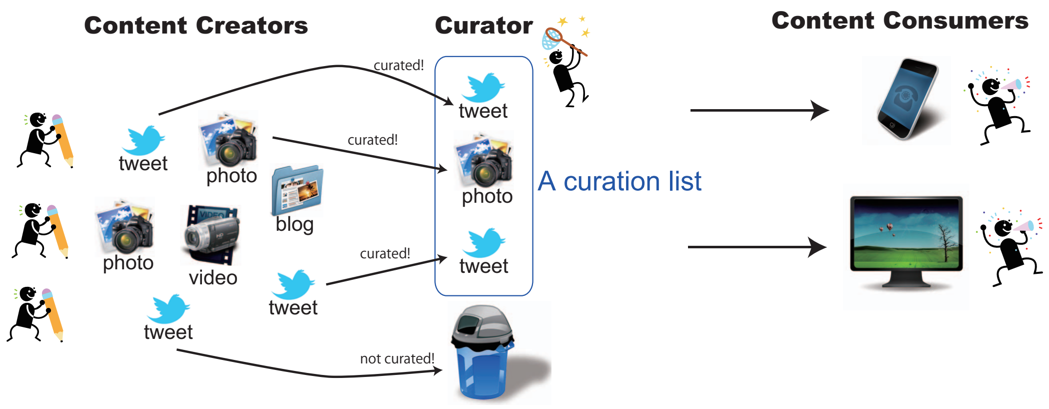


Figure 1. Schematic view of social curation service. There are three kinds of users. Content creators generate and post content in a variety of data formats and domains. Curators select posted social media content from various sources, evaluate and re-organize it as a curation list to represent their perspective, opinions, or interests. Content consumers enjoy and share these curation lists as a new type of social media content.

participants. However, the first two studies focus on the same *modality*: namely, text-based datasets. In this paper, we employ the social curation service as a complimentary information source for the automatic understanding and mining of image content in social media. This is closer to [12] in the sense that the information source is cross-modal: a social network structure with offline geographical information, as in our case social curation lists are associated with images.

To the best of our knowledge, there are no studies dealing with social curation service, excepting a work by Duh et al. [8]. This paper analyzed curation lists consisting of Twitter messages (tweets). They also studied the objectives and topics of curation lists, and reported that there are many styles and usages among social curation services. The difference from our work is again in the modality. The focus of the authors was unimodal: the authors of [8] mainly focus on text messages (i.e. tweets). In our work, we extract various kinds of information (features) from a curation list to understand and evaluate the image data.

III. Social Curation

A. Social Curation

Users involved in social curation service are classified into three types in [8] (Fig. 1). First, content creators generate *social media content* (or simply, content) that is posted to SNSs. Formats and domains of the content are diverse: text messages like tweets, photos taken by mobile phones, weblogs, movies, and so on. Second, *curators* collect and evaluate this posted content, and re-organize it to form compound content (called *a curation*, *a summary* or *a curation list*) based on the opinions, perspectives and interests of the curators. Usually, a curation list is created by one user. However, some curation lists are generated through the interaction of multiple curators. Third, content consumers enjoy, share and consume social media content created by content creators, as well content expressed by the curation lists. Note that a user can be a content creator, curator, and content consumer at the same time.

We cannot emphasize too much that each curation list is a kind of weakly supervised, organized social dataset. This means that social media items in the same curation list are expected to share the same context to a certain degree: a curation list is manually generated to fully convey one idea to the consumer. This is a very distinct characteristic compared to other social media datasets that are unorganized in many cases. Our main idea is avoiding expensive and complicated computer vision techniques for mining image content on SNSs. Instead, we employ the easily available features extracted from social curation lists for easy and precise image-content understanding. Because curation lists are weakly organized and supervised, such easy solutions may perform well against sophisticated technique applied on unorganized datasets.

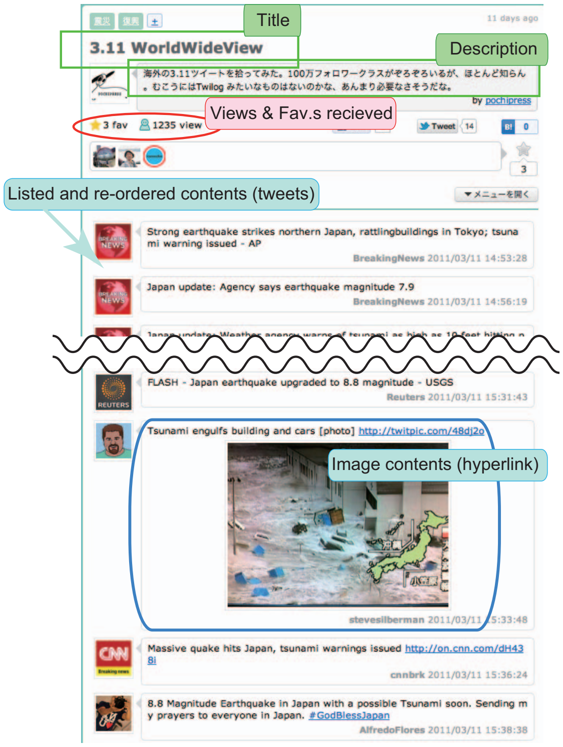
B. Dataset of Social Curation: Togetter

In this paper, we focus on the social curation service called Togetter[1]. Togetter is a social curation service mainly based on tweets (microblogs) generated on Twitter. A screen-shot of a curation list on Togetter is presented in Fig. 2. One reason for choosing Togetter is its great number of social curation lists. Togetter is rapidly growing in Japan. The number of monthly page views exceeded 10 Million by May 2012, which is three times larger than year before[2].

We collected curation lists that were created from September 2009 to April 2011 and that contain image or movie content. It is impossible to extract a complete set of such lists, thus we resort to a simple alternative. Most image and movie content is posted as a hyperlink to the file on various

[1] http://togetter.com

[2] Based on reports of donnnamedia (donnamedia.shoeisya.jp)



3.11 WorldWideView

海外の3.11ツイートを拾ってみた。100万フォロワークラスがぞろぞろいるが、ほとんど知らん。むこうにはTwilog みたいなものはないのかな、あんまり必要なさそうだな。

by pochipress

Strong earthquake strikes northern Japan, rattlingbuildings in Tokyo; tsunami warning issued - AP
BreakingNews 2011/03/11 14:53:28

Japan update: Agency says earthquake magnitude 7.9
BreakingNews 2011/03/11 14:56:19

Japan update: Weather agency warns of tsunami as high as 10 feet hitting n

FLASH - Japan earthquake upgraded to 8.8 magnitude - USGS
Reuters 2011/03/11 15:31:43

Tsunami engulfs building and cars [photo] http://twitpic.com/48dj2o
stevesilberman 2011/03/11 15:33:48

Massive quake hits Japan, tsunami warnings issued http://on.cnn.com/dH438i
cnnbrk 2011/03/11 15:36:24

8.8 Magnitude Earthquake in Japan with a possible Tsunami soon. Sending my prayers to everyone in Japan. #GodBlessJapan
AlfredoFlores 2011/03/11 15:38:38

Figure 2. Screen-shot example of a curation list on Togetter, which is a Japanese social curation service for Twitter. Tweets selected by curators are listed and displayed in an arbitrary order. Hyperlinks to other sites, and multimedia contents such as images could be included.

uploading services. We specified the major image and movie uploading services and their domains in advance. We check all tweets with hyperlinks and abbreviated hyperlinks as to whether the hyperlink matched one of these domains. A match indicated that the tweet had image or movie content.

Table I summarizes the specifications of the dataset. The number of total curation lists in Togetter was 96,506. 32,823 lists (34%) out of 96,506 lists contained an image or movie content. The number of total tweets was approximately 10 million. Of these 10 million tweets, 1.58 million contained hyperlinks to image or movie content. This means 16% of tweets in social curations refer to images or movies. Out of the 1.58 million hyperlinks, the number of unique addresses, namely unique images and movies, was 316,384.

It is surprising that the fully curated Togetter dataset contains such a large number of image and movie items. These figures indicate the popularity of image content in social media. Based on this observation, we assume that our curation dataset of Togetter is useful for confirming the automatic image understanding and mining in social media.

Table I
Specifications of crawled Togetter dataset

	Number
Total curation lists	96,506
Total curation lists including image or movie content	32,823
Total tweets	10,238,802
Total tweets including image or a movie content	1,585,448
Unique image or movie content	316,384
Unique users	106,066
Unique curators	31,661
Unique words in tweets	768,041

IV. Image View Count Prediction via Social Curation

A. Objective

Automatic image and movie understanding is one ultimate goal of computer vision research. Many researches related to visual features [13], [14], statistical models [15], [16], [17], and dataset studies [18] were studied for that purpose. These techniques enable us to recognize objects in images, detect human faces, and track moving targets in movies. However, it remains difficult to quantify subjective assessments of image content such as "*Is this image funny?*" with state-of-the-art computer vision methodology.

This paper introduces an alternative approach; we employ view counts of image content as a quantitative and measurable proxy of such evaluations. As explained, most images are stored on image uploading services. Some of those are equipped with view counters, or access counters to the contents. We think that view count is a naive but good proxy for human subjective evaluations.

Our task here is to predict view counts of image content included in curation lists based on simple features found in curation lists; no computer vision method is used. Predicting view counts of images will help catch signs of emerging trends in SNSs, and mine popular content from social media. Ideally, we would like to predict view counts of image content *before* the images are included in some curation lists. A popular image will increase views and access to the list including the image. Thus, we can recommend to curators image content that will draw much attention from content consumers. We are aware that there is some technical gap between this goal and the task validated in this paper. Thus, we focus on presenting the usefulness of social curation services and the dataset in image content mining.

B. Problem Formulation

Formally, we predict view count y_i of image or movie content i from information of the content $\boldsymbol{x}_i$. This is a typical regression problem: i.e. we try to minimize the error between the predicted view count y_i and the true view count $\hat{y}_i$ by modifying an unknown parameter $\boldsymbol{w}$ that governs the regression function $\hat{y}_i = f(\boldsymbol{x}_i; \boldsymbol{w})$.

Given image content and social curation lists, we extract several features $\boldsymbol{x}_i$ and predicted a view count for each

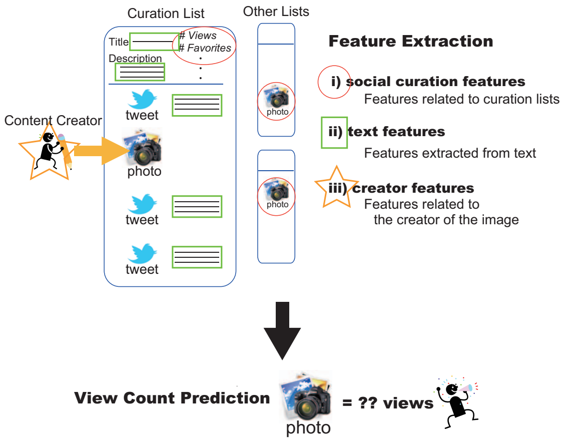


Figure 3. Experimental setup.

image content. Social curation lists contain many kinds of information that are useful for predicting view counts. For example, if the curation list that the image is included in is accessed by many content consumers, then the view count of the image content is expected to increase; or, if the image well matches the context of the curation list, the image will attract much more attention.

We compare these curation-based features to gold-standard image features used in the computer vision research community to show that social curation is a promising new data source for automatic image understanding and mining in SNSs. Note that our method is applicable to both image and movie content because it does not use any image processing.

We prepare three curation-related features in this paper as explained below.

1) Social curation features: Social curation features are computed based on social influence and characteristics of the curation list and the image content.

The first five features are dependent on the curation list the image is included in. Therefore, all images in the same list will have the same feature values.

i) The total number of tweets in the curation list.
ii) The total number of unique content creators (who post tweets) in the curation list.
iii) The total number of unique image and movie content included in the curation list.
iv) The total number of favorites the curation list receives.
v) The total number of view counts the curation list receives.

The last two statistics are officially provided by the Togetter service. These five features measure popularity and diversity of the curation list.

The following two features are dependent on the image and movie content.

i) The number of lists that contain this content.
ii) The number of tweets that contain hyperlinks to the specified image or video content.

These two features measure popularity of the image and movie content in the social curation service.

2) Text features: Text messages can directly represent the intentions, opinions, or emotions of content creators and curators. Thus, carefully designed text features would be useful in predicting responses to image and movie content. Since the objective of our problem is to compare the social curation features with computer vision techniques, however, we resort to simple and easy features. Our assumption is that if the topics or contexts of the list and the comments attached to images match well, then the images will attract much attention and gain view counts.

Our text features are computed as follows. First, we extract three parts of texts that are found in Togetter curation lists, and compute a Bag-of-Words (BoW) histogram from each part. The first part is the title and description of the curation list. This part is edited directly by curators; so we expect this part concisely describes the entire context of the list. The second part is all texts of the curated contents in the list: tweets, comments to images, and so on. The BoW histogram of this part is a direct summarization of the curation list. The third part is all tweets among curation lists that have a hyperlink to the image content. The histogram of this part encompasses the responses of SNS users with regard to the content. From these BoW histograms, we compute three cosine distances based on our assumption.

i) Distances between the first and the second BoWs.
ii) Distances between the first and the third BoWs.
iii) Distances between the second and the third BoWs.

Feature i) computes text context similarities between the {title, description} and the tweets in the list. In other words, this feature is a measure of the similarity between curators' intention and the actual context of the list. Feature ii) computes text context similarities between the {title, description} and the responses to the focused image content. Namely, this feature is a measure of the similarity between curators' intention and observed responses to the image content in SNSs. Feature iii) computes text context similarities between the tweets in the list and the responses to the image content. In other words, this feature is a measure of the similarity between the actual context of the list and observed responses to the image content in SNSs.

Also, we compute binarized versions of three BoWs. We binaraize the BoW histograms by thresholding in order to absorb the difference in lengths and numbers of tweets among the curation lists. We compute three cosine distances for these binarized BoWs in the same manner. Thus, we finally obtain six text features.

3) Content creator features: We observe that there might be a correlation between the view count of an image content and the information provided by the creator of that content.

We compute the following features related to creators of the content to utilize this correlation.

i) The number of users who follow the creator of the content.
ii) The number of users the creator of the content follows.
iii) The number of favorites set by the creator.
iv) The number of "Lists" (an official Twitter function, not curation lists) the creator is included in.
v) The main language of the creator of the content.
vi) When the creator of the content started using Twitter.

4) Combining features: We concatenate three types of features into a single vector and use the vector as the social curation feature vector of content i $\boldsymbol{x}_i^S$.

There are several approaches to combining the three types of features. In the experiments, we tested the following combinations: i) social curation features, ii) social curation features + text features, iii) social curation features + creator features, and iv) social curation features + text features + creator features. If an image is selected by multiple curation lists, we simply add the feature values taken from all lists.

5) Regression model: We employ Support Vector Regression (SVR) (cf. [19]) as the regression function. SVR is known for its powerful regression performances, and is used as one of the standard regression models. We use an implementation provided by libSVM [20]. As the kernel function, we choose the standard RBF kernel. We experimentally optimized the soft margin parameter and the kernel parameter. Other parameters were set to default values.

V. Experiments

A. Image features and data preparations

For comparisons, we need to choose visual features for image content. In this paper, we focused on only image content for ease of experimentation, and chose SIFT [13] as the main visual feature: a gold-standard feature in current computer vision research. A SIFT feature is defined in a local "interest" point in the image; it measures the orientations of edges at multiple scales. We collect a set of SIFT features from many interest points of among images, and discretize them by K-means clustering. After that, the set of SIFT features of an image is encoded to a histogram of Bag-of-visual words (K words) as in BoW of texts.

We employ an implementation provided by the author of [21] for image feature extraction. We extract SIFT, C-SIFT [22], OpponentSIFT, and Transformed Color Histogram features from each image. All features are discretized into K visual words, and encoded as histograms of K visual words. We have tested two types of image feature usage. The first one only employs SIFT features, resulting in K-dimensional image feature vector $\boldsymbol{x}_i^I$. The second one employs all four features, resulting in $4K$-dimensional $\boldsymbol{x}_i^I$.

Next, we explain how we prepared the dataset. Since we focus on only image content, we cannot utilize the whole dataset of Togetter. Also, the images used in the experiments must be linked to true view counts. We found 22,024 images satisfying this requirement, and used all of them in the experiments. The vocabulary size of text features was set to $V = 50,000$. The vocabulary size of image features was set to $K = 1,000$. The evaluation criterion was mean squared error (MSE). We computed the MSEs of each feature by 10-fold cross validation.

The distribution of view counts is skewed: the minimum view count is 0, while the maximum count is 1,288,507. Therefore, we used the logarithm of view counts in the experiment. This yielded the average and the variance of the log view counts of 4.3698 and 3.0125, respectively.

B. Results and Discussions

Table II lists the MSEs for each feature choice. As evident from the table, social features $\boldsymbol{x}_i^S$ always beat image features $\boldsymbol{x}_i^I$. The MSEs of social features are almost half those of image features, which are high-dimensional and known for great performance in computer vision tasks. In fact, the MSEs of image features are close to the variance of log view counts. This means that the predictions by image features are at the "chance level".

As the table shows, a combination of social curation features and text features improves the prediction accuracy slightly. However, incorporating text features to a combination of social curation features and creator features degrades the MSE. For the best performance, obviously we need further investigation of feature designs and combinations.

libSVM [20] provides a feature scaling function in order to absorb the scale differences among feature elements. We re-scaled all the feature elements between $[0, 1]$, and reran the experiments. The results are also shown in Table II. We confirmed that the social features are much better than image features. Also, we observed that feature scaling improves prediction accuracy in some cases, but not in others. To further improve prediction accuracy, we need to more carefully consider the differences in scales between features.

Finally, we discuss the differences between our results and the works by van Zwol et al. [23]. The authors of [23] proposed a classifier that predicts whether a user will favor a photo on Flickr. They tested combinations of text features, visual features and social features, and reported that the social features performed well, and combinations with other features improve, in most cases, prediction accuracy. This is different from our results. One possible explanation is that popularity among SNSs differs fundamentally from the preference of a specific user: in SNS, popularity is the result of many users with different preferences.

VI. Conclusion

In this paper, we focused on the emerging *social curation* service as a new source for image and movie mining in social

Table II
Mean squared errors (MSE) of view count regression by SVR. The target is a natural logarithm of view counts. Smaller values are better.

Types of features	Feature Dim.	MSE	MSE(scaled)
Social curation	7	1.5441	1.4247
Social curation + text	13	1.5234	1.9646
Social curation + creator	13	**1.2650**	**1.1629**
Social curation + text + creator	19	1.2668	1.3013
image features (SIFT)	1000	3.0289	3.3190
image features (All)	4000	3.0287	3.3124
var. of log view counts	-	3.013	3.013

media. A key insight is that a curation list, which is a user-generated agglomeration of social media content, is weakly supervised: *represent the manual collection, selection, and maintenance by curators*. This is a unique characteristic compared to other social data, and we are able to understand and mine images in SNSs by fully utilizing social curation data. We confirmed that our Togetter dataset contained many image and video items. Experiments were conducted on the prediction of view counts of image content as a first step for automatic image understanding and mining in social media. The results show that the social curation information is far superior in predicting the view counts of images to the gold-standard image features used in computer vision research.

In this paper, we investigated only a specific curation dataset for a specific task. We are aware that there are many open problems. First, there is no guideline for designing and choosing social curation features, and combining visual features. Second, we have to investigate social features in a larger dataset, and other tasks such as image retrieval. Finally, applying social curation information to other domains such as natural language processing, music and audio processing would be fruitful for further development of social data mining technologies.

References

[1] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in *Proc. WWW*, 2011.

[2] H. Lakkaraju, A. Rai, and S. Merugu, "Smart news feeds for social networks using scalable joint latent factor models," in *Proc. WWW*, 2011.

[3] J. Lin, R. Snow, and W. Morgan, "Smoothing techniques for adaptive online language models: Topic tracking in tweet streams," in *Proc. KDD*, 2011.

[4] D. Rao, M. Paul, C. Fink, D. Yarowsky, T. Oates, and G. Coppersmith, "Hierarchical Bayesian models for latent attribute detection in social media," in *Proc. ICWSM*, 2011.

[5] Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and llighting," in *Proc. CVPR*, 2004.

[6] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic object recognition with boosting," *IEEE Trans. PAMI*, vol. 28, no. 3, pp. 416–431, 2006.

[7] D. Parikh and K. Grauman, "Interactively building a discriminative vocabulary of nameable attributes," in *Proc. CVPR*, 2011.

[8] K. Duh, T. Hirao, A. Kimura, K. Ishiguro, T. Iwata, and C.-M. Au Yeung, "Creating stories: Social curation on twitter messages," in *Proc. ICWSM*, 2012.

[9] Y. Mejova and P. Srinivasan, "Crossing media streams with sentiment: Domain adaptation in blogs, reviews and Twitter," in *Proc. ICWSM*, 2012.

[10] Y. Hu, A. John, D. D. Seligmann, and F. Wang, "What were the tweets about? topical associations between public events and twitter feeds," in *Proc. ICWSM*, 2012.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[12] J. Kulshrestha, F. Kooti, A. Nikravesh, and K. P. Gummadi, "Geographic dissection of the twitter network," in *Proc. ICWSM*, 2012.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[14] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," in *Proc. ICCV*, 2003.

[15] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. CVPR*, 2001.

[17] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros, "Unsupervised discovery of visual object class hierarchies," in *Proc. CVPR*, 2008.

[18] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large dataset for nonparametric object and scene recognition," *IEEE Trans. PAMI*, vol. 30, no. 11, pp. 1958–1970, 2008.

[19] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[20] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[21] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. PAMI*, vol. 32, no. 9, pp. 1582–1596, 2010.

[22] G. J. Burghouts and J.-M. Geusebroek, "Performance evaluation of local colour invariants," *Computer Vision and Image Understanding*, vol. 113, no. 1, pp. 48–62, 2009.

[23] R. Van Zwol, A. Rae, and L. G. Pueyo, "Prediction of favourite photos using social, visual, and textual signals," in *Proc. ACMMM*, 2010.