Context-aware Image Tweet Modelling and Recommendation

Tao Chen[1,2] Xiangnan He[1] Min-Yen Kan[1,2]
[1]School of Computing, National University of Singapore
[2]NUS Interactive and Digital Media Institute
{taochen, xiangnan, kanmy}@comp.nus.edu.sg

ABSTRACT

While efforts have been made on bridging the semantic gap in image understanding, the *in situ* understanding of social media images is arguably more important but has had less progress. In this work, we enrich the representation of images in *image tweets* by considering their social context. We argue that in the microblog context, traditional image features, *e.g.*, low-level SIFT or high-level detected objects, are far from adequate in interpreting the necessary semantics latent in image tweets.

To bridge this gap, we move from the images' pixels to their context and propose a **c**ontext-aware **i**mage **t**weet modell**ing** (CITING) framework to mine and fuse contextual text to model such social media images' semantics. We start with tweet's intrinsic contexts, namely, 1) text within the image itself and 2) its accompanying text; and then we turn to the extrinsic contexts: 3) the external web page linked to by the tweet's embedded URL, and 4) the Web as a whole. These contexts can be leveraged to benefit many fundamental applications. To demonstrate the effectiveness our framework, we focus on the task of personalized image tweet recommendation, developing a feature-aware matrix factorization framework that encodes the contexts as a part of user interest modelling. Extensive experiments on a large Twitter dataset show that our proposed method significantly improves performance. Finally, to spur future studies, we have released both the code of our recommendation model and our image tweet dataset.

Keywords

image tweets; Twitter; context; recommendation; image semantics; microblog

1. INTRODUCTION

In the mobile Internet era, people now effortlessly snap pictures and share the events in their daily lives on social media. As a result, usage of social media platforms has soared, especially in terms of user-generated images. For example in Twitter, about 17.2% of tweets have associated images (which we term **image tweets**) according to our dataset collected in December 2014 (detailed in Section 5). The trend is even more evident in Sina Weibo — the largest microblog service in China — where over 45% of tweets are image tweets [8]. Being able to understand the images in image tweets is beneficial for many downstream applications, such as event detection, image tweet retrieval and recommendation [35]. For example in tweet recommendation, the widely-used collaborative filtering technique does not work well [7, 10, 15], due to the short life span of tweets. Therefore, it is crucial to look into the content of image tweets.

Figure 1: Example images affiliated with two tweets: (left) China ends the one-child policy, and (right) the movie *Fast and Furious 6*. For such microblog images, understanding their context is the key to its semantics.

Understanding image tweets, however, is non-trivial for machines. This is partially due to the technical difficulties in mining semantics from images, but more importantly, is because of the special **context** in social media. A picture is worth a thousand words, while the ever-evolving property of social media induces the full story of an image is far beyond itself and the contexts (*e.g.*, event and intent) are critical for image tweets understanding. Figure 1 shows two examples. For the left picture, image recognition algorithms tag it with words like "child, cute, girl, little, indoor"[1]. However, these visual tags can not capture the background and real intent of the picture — this image was the poster child for the story of China abandoning its controversial one-child policy. Similarly, for the right picture, the annotated tags "car, asphalt, road, people, transportation system" fail to tell the origin and objective of the picture — this is a promotional poster for the movie *Fast and Furious 6*.

In this work, we propose to exploit the contexts to tackle the challenges in understanding image tweets. We devise a **c**ontext-aware **i**mage **t**weet modell**ing** (CITING) framework (illustrated in Figure 2) to enrich the representation of im-

[1]These are the actual top five tags from Clarifai (https://www.clarifai.com), a commercial visual recognition system.

MM '16, October 15–19, 2016, Amsterdam, The Netherlands.

DOI: http://dx.doi.org/10.1145/2964284.2964291

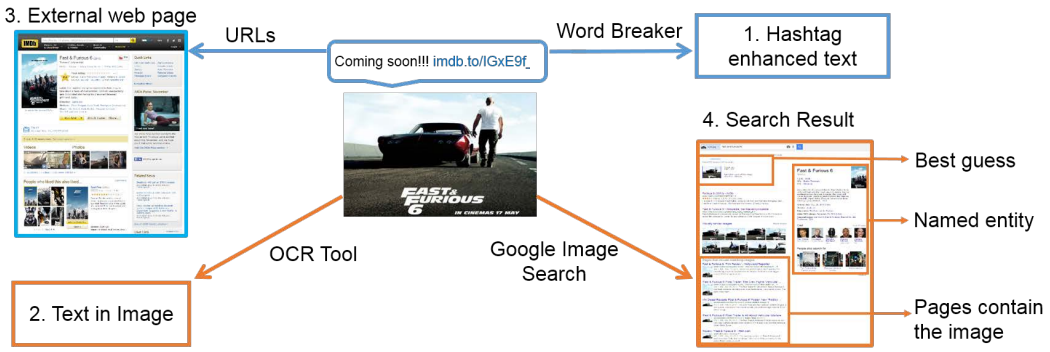

Figure 2: An image tweet's four sources of contextual text in our CITING framework. Blue outlines denote evidence frome text; orange from the image.

age tweets from both intrinsic and extrinsic contexts. We start with the intrinsic contexts: 1) for the text, we enhance hashtags to better represent the topics of images, and 2) for the image, we apply optical character recognition (OCR) to extract text from images. Then we turn to extrinsic contexts, which are especially important to understand the full story of image tweets: 3) parsing the external webpage(s) hyperlinked within the tweet; and 4) querying image search engines with the image as query. Our assessment reveals that the contextual text from each source differs in quality and coverage. As such, we further devise a series of heuristics to fuse text when multiple channels are triggered. This fusion makes the modelling more accurate and also reduces the acquisition cost of the contexts.

Our proposed framework extracts contexts in the form of textual words, which are easy to interpret and integrate with other image/visual feature inventories. To demonstrate the efficacy of the extracted contexts, we apply them to the personalized image tweet recommendation task, for which the key is to accurately model users' interests. We develop a generic feature-aware Matrix Factorization (MF) framework to model users' preference on features. As users do not explicitly express their dislikes, there is no explict negative data, which can adversely affect the learning of user interests [13]. To resolve this, we propose a time-aware negative sampling strategy that samples negative tweets for a user based on how likely the user may see the tweet but has not retweeted it. Lastly, we adopt a pair-wise learning to ranking method to infer users' interests based on our enhanced contexts. We conduct experiments on a large Twitter dataset, showing that our proposed contexts are more effective for users' interests modelling than the tweet's text and visual images themselves, which validates our recommendation methodology.

We summarize the main contributions as follows:

- We propose the CITING framework to mine and fuse contexts to better represent image tweets. Such study is fundamental and can benefit many applications such as event detection, image tweet retrieval and recommendation (Section 3).
- We develop a feature-aware recommendation model together with a featured negative sampling method for tweets, so as to effectively integrate contextual features for user interest modelling (Section 4).
- We conduct extensive experiments on Twitter image tweets that show the effectiveness of our CITING framework and the recommendation model. We have made the dataset and the model publicly available to facilitate future research (Section 5).

2. RELATED WORK

Unlike their textual counterparts, images in microblogs have only started attracting academic attention recently. We review the existing studies on image tweets, focusing on how the semantics of images are exploited for downstream applications. As there is no previous work on personalized recommendation of image tweets (to the best of our knowledge), we then review works about general tweet recommendation in the microblog setting.

2.1 Images in Microblog

As an image tweet consists of both text and image components, most existing works leverage both modalities to interpret the semantics of an image tweet [2, 3, 6, 30, 31, 34]. For text part, it is widely accepted that pre-processing — such as tokenization, lowercasing and stopword removal — is important to combat noise in the text. For the image part, existing work has followed a multimedia paradigm; that is, attempting to mine the semantics of an image from both low-level and high-level features. In terms of low-level features, works [30, 31, 16, 2, 3, 26] have considered pixels, color histograms, SIFT descriptors and Speeded Up Robust Features (SURF), among others. To be specific, SIFT descriptors and SURF are quantized by means of a visual codebook learned by k-means. However, the gap between the low-level features and the real semantics limits the model fidelity. The other works leveraged higher-level features, such as visual objects [6], human faces [8], and the output from the upper layers of convolutional neural networks (CNNs) trained for object recognition (hereafter, CNN features) [4, 34, 5]. For the CNN features, two papers [4, 5] used them directly (*e.g.*, 4096 dimensional visual features), while Zhang *et al.* [34] carried additional steps to quantize them into discrete features, which shares a similar spirit to the quantization of SIFT and SURF.

Due to the heterogeneous nature of image tweets, features extracted from the text and image lie in different spaces, which are non-trivial to integrate. To resolve this, multimodal topic models [30, 31, 2, 4, 34, 9] have been developed to project images and the text of tweets to a shared latent topic space. As such, an image tweet is represented as a

multinomial topic distribution; however, the dimensions of such latent representation are difficult to interpret semantically. One solution is to assign labels or categories to image tweets through human labelling or transfer learning [3]. However, the limited number of labels can seriously restrict the method's generalizability and coverage in practice. For example, in Bian *et al*'s work [3], only 20 categories were considered to label image tweets, which significantly limited the representation of semantics.

Although the above works have exploited image tweets in various ways, we argue that there are two key limitations of the previous work. First, concerning the tweet's text, only the textual words have been investigated; but microblog-specific textual features (*e.g.*, hashtags, external URLs) that contain known, rich contextual information have not been utilized. Second, images are leveraged only at a shallow level; that is, the community has limited itself to only using image features commonly used in general domain research. Microblog images exhibit unique characteristics, such that modelling these images using conventional manners confines an understanding to a superficial level (see the examples in Figure 1). It is thus critical to turn to external knowledge to gain a comprehensive understanding of an image tweet. In this work, we aim to fill this research gap by mining the rich contexts of image tweets.

2.2 Tweet Recommendation

With the vast amount of tweets, microblog users are now overwhelmed with many uninteresting posts. It is of great necessity to understand users' interest and recommend interesting tweet feeds for users. One line of research [14, 23, 22, 29] attempted to predict the general interestingness (or equivalently, "popularity") of a tweet, in regardless of the identity of an audience. Such prediction task is usually formulated as a classification problem (*e.g.*, popular or not). To this end, various features have been exploited, such as the explicit features from tweet's textual content (*e.g.*, words, topics and sentiments), contextual meta-data (*e.g.*, posting time), and the author's profile (*e.g.*, the number of followers and followees). The only work that has paid attention to image tweets is done by Can *et al.* [6], which also utilized shallow image features to build a classifier.

However, a generally popular tweet does not necessarily mean it will be interesting to a particular user, since interestingness is subjective and relevant to user's own taste [1]. To generate better recommendations for users, researchers have turned to build personalized models to predict tweet's interestingness. An early work by [28] built a classifier similar to general popularity predictor but with additional features from the target user, such as user's retweeting regularity and user–author relations. Later work casted it as a typical recommendation problem [7, 15, 10, 33], for which collaborative filtering (CF) is known to be the most effective technique. In the microblog platform, however, CF does not work well for tweet recommendation because of the ubiquitous cold-start problem: most live tweets are newly generated and have never been seen in the training data. To tackle this, existing works incorporate tweet's textual content into collaborative filtering models. Specifically, Chen *et al.* [7] transformed the traditional user–tweet interaction matrix to a user–word matrix before applying matrix factorization. Following the same idea, Feng *et al.* [10] additionally modelled the user–hashtag interaction, since hashtags are good topic indicators. Hong *et al.* [15] extended the Factorization Machines to jointly model user–tweet relations and the textual tweet generation process.

Despite the fact that many works have studied the tweet recommendation problem, they have primarily focused on the textual tweets. The rich signals in images and their contexts have been ignored. To the best of our knowledge, our work is the first to specifically consider the personalized recommendation task with image tweets.

3. CONTEXT-AWARE IMAGE TWEET MODELLING

In this section, we present our CITING framework for image tweets modelling. We first describe the four strategies that construct contexts from different data sources (*cf.* Figure 2), and then discuss the rules to fuse the contexts which help to improve text quality and save the acquisition cost. Unless otherwise stated, the descriptive statistics in this section are drawn from our 1.3 million Twitter image tweets dataset (detailed in Section 5).

3.1 Four Strategies to Construct Contexts

We start with the intrinsic context in image tweets: 1) the textual tweet, and 2) the image itself. Then we turn to the extrinsic context: 3) the external web pages hyperlinked in the tweet, and 4) the whole Web based on search engine.

1. Hashtag Enhanced Text. The most obvious context for a microblog image is its accompanying text. Here we focus on hashtags, which have relatively high coverage due to their prevalence in image tweets — 26.8% have hashtags in our Twitter dataset. Compared to the textual words of a tweet, hashtags exhibit stronger semantic ties to the post [21]. For example, we observe that a few hashtags (*e.g.*, #dogphoto) annotate objects present in an image, while the majority describe the topic or event of the image (*e.g.*, #itsyourbirthday). In both cases, hashtags are helpful in capturing the semantics of the image. However, a challenge in utilizing hashtags is that they do not exhibit the regularity of controlled vocabulary due to the user-generated nature. More specifically, people usually use different variant hashtags to refer to the same (series of) events (*e.g.*, #icebucket, #ALSIceBucketChallenge; #NewYears2013, #NewYears2014). Gathering hashtags variants can thus help conflate images with common semantics. Observing these variants are often composed with common keywords, we break up hashtags into component words by Microsoft's Word Breaker API[2], *e.g.,* #icebucket will be broken up as "ice" and "bucket"; we found that 14.3% of image tweets utilize multiword hashtags. We then combine such component words with post's text (including the original hashtags with their hash symbol) to form the hashtag enhanced text.

2. Text in the Image. Images in microblogs are not solely captured by camera, and many of them are software-generated or edited images, *e.g.*, graphics, memes, cartoons and screenshots. We observe that text is often embedded in images: our own manual annotation[3] of 500 randomly-sampled images (hereafter, *Twitter-500*) from our Twitter dataset identified 174 (34.8%) that fall in this category. We

[2]https://www.projectoxford.ai/weblm

[3]Annotated by the first author.

Table 1: Demographics of the 5 subtypes of text-images and associated Tesseract OCR performance.

Category	Manual	Tesseract	
	# (%)	Miss Rate	Recall
Text-style	38 (21.8%)	10.5%	0.984
Meme-style	64 (36.8%)	42.1%	0.572
Tweet screenshot	14 (8.0%)	7.1%	0.843
Other synthetic	43 (24.7%)	30.2%	0.500
Natural scene w/ text	15 (8.6%)	66.7%	0.467
Total	174	119	

Table 2: Categories of the 100 most frequent domains in URLs and Google Image indexed pages. *YT* is YouTube and *IA* is image aggregator. For the 66.0% SNS indexed by Google, 48.0% are from Twitter, 40.1% are from Pinterest.

%	News	SNS	Shop	Article	YT	IA	Music
URLs	**51.7**	15.3	11.9	10.9	3.9	2.3	1.8
Google	5.7	**66.0**	3.6	0.3	2.8	18.3	1.0

term such images as *text-images*, which we further categorize into five subtypes as shown in Table 1.

First, from the second column, we see that one-third of text-images are meme-styled: *i.e.*, a (viral Internet) image overlaid with text (as in Figure 3, left). It is impossible to differentiate the semantics of meme-style images from a visual perspective, as many originate from an identical source picture. Figure 4 shows two example images. In contrast, the embedded captions all but give away the context. In even more text-heavy cases, images can consist purely of text (Figure 3, right), accounting for roughly a fifth of text-images. Twitter users sometimes post such pure text-style images possibly to circumvent the 140 character restriction. Screenshots of tweets (8.0%) are also common; we conjecture that the primary intention of such posts is to achieve the "retweet with comment" feature before Twitter officially supported this function in April 2015[4]. For such tweets that have a strong textual nature, object detectors are close to useless. For the remaining text-images, 16.7% are other synthetic images, and 8.5% are natural photos that contain text in the scene (*e.g.*, road signs). Our findings lead to two key implications: 1) that a large proportion of social media images have a textual aspect, and for posts feature in such images, that 2) the embedded text is an important carrier of its semantics.

As such, we apply the *Tesseract* open source OCR software (version 3.02.02)[5] to recognize text from these images. After further using the vocabulary built by our Twitter posts to filter out noise, 26.4% of the images in our dataset have at least one recognized textual word. As Tesseract is designed for printed text, a natural question to ask is how well does it work for Twitter images? Using our manual annotation as a reference, we evaluate Tesseract's performance on our Twitter-500 sample set. Table 1's rightmost two columns show its miss rate and the average recall for recognized text. Overall, Tesseract detected text from 119 images, missing 55 images that actually did contain some text. The majority of the misses come from text present in the scene (missed

[4]It is likely that the number of such screenshots is decreasing, while we believe the overall coverage of text-images has not change much.

[5]https://github.com/tesseract-ocr/tesseract

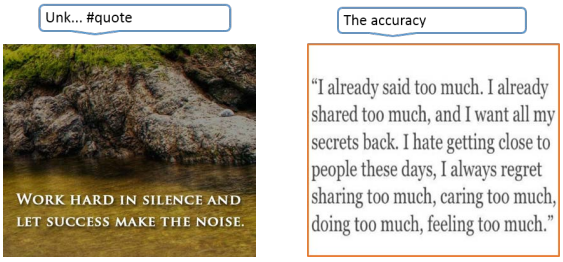

Figure 3: (left) Meme-styled and (right) text-styled image tweets. The callouts are the tweets' text.

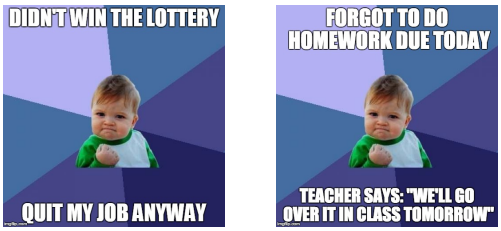

Figure 4: Two meme-styled images have similar visual properties but different embedded captions.

two-thirds) and meme-style text (missed 42.1%). Tesseract performs well on pure-text style images (detected 89.5% of images with some text, and recognized 98.4% of the actual words) and tweet screenshots. The cause of the discrepancy is simple: the more similar the image is to scanned text, the better the performance.

3. External Webpages. To provide context as well as to circumvent length limitations, microblog users also embed shortened hyperlinks in their tweets. In our dataset, 22.7% of image tweets contain at least one external URL. To the best of our knowledge, URLs in image tweets have not been studied in prior work. What are the external web pages about? How do they correlate to the images?

To answer these questions, we first resolved the hyperlinked shortened URLs and stored the redirected original URLs[6]. We then aggregated the resolved URLs by domain, manually categorizing the top 100 most frequent domains (accounting for 51.8% of URLs) into seven types. Table 2 shows the category distribution of the external resources. The majority are news reports, while three other prominent sources are online social networks (15.3%), e-commerce shops (11.9%), and articles (10.9%, *e.g.*, WordPress blogs). YouTube, image aggregators and music links account for the remaining minority (3.9%, 2.3% and 1.8%, respectively).

Interestingly, we also discover that the tweet image often originates from the external resource (82.1% of URL image tweets in our Twitter-500 set). Often, the image is a key scene in a news event, an item to be sold for online shops, or a portrait of the musician in music links. This suggests the external resource is the original, unsummarized context for such tweets, and thus a reliable source for capturing the image's semantics. We thus apply Boilerpipe [18] to extract the main textual content, then filter out stopwords, and finally use standard $tf.idf$ term weighting to select the top k textual words as features. Considering some pages consist only of the title text (no main text), we use the page's title as another descriptor.

[6]Over half were still accessible, as of 30 September 2015.

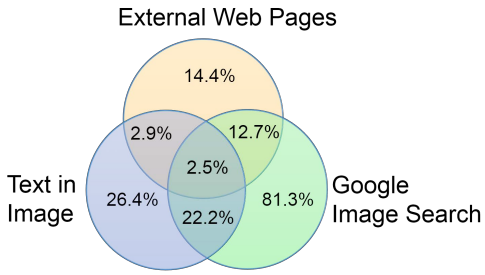

Figure 5: Percentage of image tweets that benefit from three major sources and the overlaps.

4. Search Engine as Context Miner. As 85% of Twitter trending topics are news [20] and Internet viral images are popular, many such tweet images have been previously used in other places on the Web, in similar contexts. To obtain these external contexts, we leverage Web search engines, which represent an up-to-date repository for the Web. We send each image in our dataset as a query to Google Image Search (done during the last week of August 2015), then parse the first search engine result page (SERP) to obtain a list of pages that contain the image (including URL and title). We then follow the links to crawl the actual content of the external pages. In our dataset, a surprisingly large proportion (76.0%) of Twitter images were already been indexed by Google.

Following our workflow for tweets' embedded URLs, we also categorize the top 100 domains for such SERP-listed web pages, which accounted for 54.6% of pages. From Table 2, we see 66.0% of pages are social network posts, of which 48.0% originate from Twitter itself. This implies images are re-purposed even in Twitter, and that image reuse is not limited to retweeting. The photo-based Pinterest social network takes up another 40.1% of such posts. The second largest category represents photo aggregators (18.3%, *e.g.*, imgur.com), which host images for social networks. The remaining 15.7% is distributed among the other site types (news sites, e-commerce, YouTube, music sites and blog sites, representing 5.7%, 3.6%, 2.8%, 1.0% and 0.3%, respectively).

For the query image, Google Image Search occasionally also offers a "best guess" at a short text description. Unlike the tags from traditional visual recognizer, the "best guess" can be seen as translating from a visual description to a semantic description — technically implemented as the best keyword for discovering the query image. For Figure 1 (right), the best guess is "fast and furious 6" which is spot-on. When the query image is identified as a named entity (*e.g.*, celebrity, movie or landmark), Google also sometimes shows a detailed named entity description in a knowledge graph box (functionality introduced in Google around July 2012). We additionally utilize these sources — the best guess (57.9% of Twitter images) and named entity (8.1%) as image's semantic description when available. In sum, 81.3% of images in our dataset have obtainable contextual text from Google Image Search.

3.2 Fusing the Contexts

Image tweets have rich contexts that can be exploited. In our dataset, 89.1% of images have at least one applicable strategy and 39.9% can leverage multiple ones. We

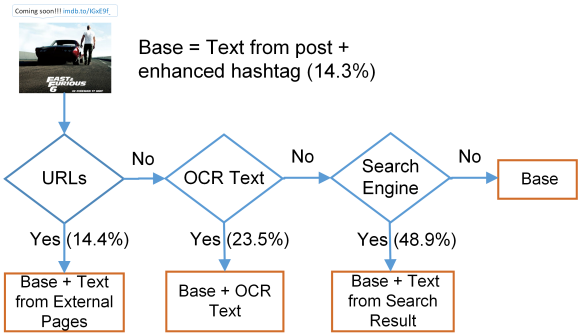

Figure 6: Our filtered rule for fusing text from context sources. The % denotes the coverage of each source alone after fusion.

survey the overlap among the contextual text sources of external URLs, OCR-ed text and Google Image Search for our dataset in Figure 5. As we can see, Google Image Search has large overlaps with external URLs and OCR text. For these overlaps, the other two sources are direct context indicated by image tweet's author, and we believe provide more accurate semantics for the image tweet than the SERP-extracted text. Take the two meme-styled images in Figure 4 as an example. The best guess description from Google Image Search is "no adulting meme" and "india pakistan match troll", respectively. Neither reveals the correct semantics which the OCR text does.

As such, instead of merely polling all four sources of contextual text, we can fuse them more opportunistically to improve text quality. We feel that this cascading approach is a better option than a weighted sum, as it also cuts down computation costs. The tweet's original textual post and enhanced hashtags form the basis for fusion, as they are the most obvious context created directly by the author. We then propose a filtered fusion approach (illustrated in Figure 6) to use text obtainable from the other three sources: 1) for an image tweet with an embedded URL, we fuse only the text from its external web page, since the external page is the most accurate and accessible semantic context for the image; 2) for the remainder, we apply OCR on the image and if it contains embedded text, we fuse its OCR text recognized by Tesseract; 3) but if no embedded text is found, we obtain and fuse the SERP-extracted text from Google Image Search. It is worth noting that the fusion strategy helps to reduce the acquisition cost of contexts by 18.0% in our dataset (when treating all API calls as a unit cost), and provides better semantic modelling for image tweets (demonstrated in Section 5.2).

4. PERSONALIZED IMAGE TWEET RECOMMENDATION

We now apply the CITING contexts that encode image tweets' semantics for personalized image tweet recommendation. To the best of our knowledge, this is the first study that learns user's interest from image tweets. To be specific, for a particular user, we aim to model her interest from her previous history, and predict her interest in incoming new image tweets. A direct application is to reorder the image tweets in user's feeds according to their interestingness.

We first discuss traditional collaborative filtering tech-

Figure 7: An example of user–item matrix, where 1 denotes the user has retweeted the image tweet and 0 otherwise. The rightmost two columns denote two new items that cause the cold-start problem.

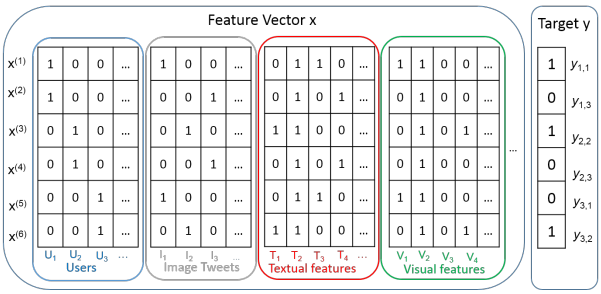

Figure 8: Example feature vectors for the user–item matrix in Figure 7. Each row consists of user ID, item ID, and the various features of the item. The rightmost column is the prediction target.

niques for tweet recommendation, then detail our proposed feature-aware Matrix Factorization model, a generic method that can incorporate various features for image tweet recommendation.

4.1 Drawbacks of Collaborative Filtering

Collaborative filtering (CF) is acknowledged to be the most effective and generalizable technique for recommender system [25]. The basic idea is to predict a user's preference based on the histories of other similar users. For example, Matrix Factorization (MF) [19], the most popular CF method, projects users and items into latent space to encode the preference similarity. As CF is designed to operate on the user–item interaction matrix, it represents an item as an ID and thus learns user's preference on item IDs.

We highlight that a key weakness of CF is its inapplicability to new items that have not yet attracted any interaction, which is known as the *cold start* problem. We show an illustrative example in Figure 7, where the rightmost two columns denote two new image tweets, which have never been seen in the training set. In this case, CF will fail to infer users' interest on the new items, and the prediction on new items is no better than random. This phenomenon is even exacerbated in social media like Twitter, due to the medium's strong timeliness and dynamicity. Said differently, old tweets (that are in training set) can quickly become dated and unattractive, while new tweets can be interesting but never appear in training. This makes the traditional CF technique unsuitable for the tweet recommendation task.

4.2 Feature-aware MF Framework

To overcome the defect of CF, a common solution is to go beyond modelling the interaction of user and item IDs which are sparse for new items, and to model more dense interactions by examining the properties of the content itself as features. As a consequence, although the ID of a new item has not been seen before, we can still infer a user's preference on new items based on their features (whose preferences have been learned during training). This motivates us to develop a generic model that can capture the interaction with various features for recommendation.

Following the paradigm of factorization machines (FM) [24], in our framework, we transform the user–item interaction matrix to a set of feature vectors (model input) and a vector of interactions (target). As illustrated in Figure 8, each row denotes an interaction consisting of user ID, item ID, and the various types of contextual features of the item. In our feature-aware MF model, we learn a low-dimensional representation (also termed as "latent vector") for each user and feature. Suppose we have N types of features that represent image tweets, then our model estimates the preference of user u on image tweet i as:

$$\hat{y}_{u,i} = \mathbf{v}_u^T \left(\sum_{n=1}^{N} \frac{1}{Z_{n,i}} \sum_{f \in \mathbb{F}_{n,i}} \mathbf{q}_f\right), \tag{1}$$

where $\mathbf{v_u}$ and $\mathbf{q_f}$ denote the latent vector for user u and feature f, respectively. $\mathbb{F}_{n,i}$ denotes the feature set of the n^{th} feature type for item i, and $Z_{n,i}$ is the normalizer for features (setting $Z_i = \sqrt{|\mathbb{F}_{n,i}|}$ leads to good results).

This model is generic that it can incorporate any type of features (and their combinations). In the case of image tweets, they can be image features (*e.g.*, visual objects), textual words in tweets and our proposed contexts[7]. We hypothesize that incorporating our proposed contextual features will better capture the rich semantics in image tweets, leading to better personalized tweet recommendation.

We point out that the key difference with FM is in the feature interactions considered — FM models the interactions between all pairs of features (including feature pairs of the same feature type), while we only model the interactions between user and item's features. Our design choice is for the model's interpretability. By modelling only the interaction between user and feature, we can interpret user u's preference on feature f as the inner product $\mathbf{v}_u^T \mathbf{q}_f$, which benefits the explainability [12] of our recommendation model.

4.3 Learning from Implicit Feedback

The objective of tweet recommendation is to provide a user with a personalized ranked list of tweets. From a user's observed behaviors (*e.g.,* retweets), we naturally have the positive feedback that represents which tweets users are interested in. These positive tweets should be ranked higher than other negative tweets for the user. Since the idea naturally fits the framework of pairwise Learning to Rank, we adopt the *Bayesian Personalized Ranking* [25], which learns the model to maximize the probability that the positives should be given a higher score than the negatives for users:

$$p(.|\mathbf{V}, \mathbf{Q}) = \prod_{u \in \mathbb{U}} \prod_{i \in \mathbb{P}_u} \prod_{j \notin \mathbb{P}_u} \sigma(\hat{y}_{u,i} - \hat{y}_{u,j}), \tag{2}$$

[7]Note that we do not insert item IDs into the feature vector (in gray in Figure 8), since most items in test set are not observed in training. Excluding item IDs lessens the bias to recommend only tweets seem in training, and favors the prediction of new cold-start items, leading to better results.

where $\mathbb{P}_u$ denotes the positive tweets for user u, and σ is the sigmoid function that projects the margin value into probability space. We note that aside from the pairwise function, another option for learning from implicit feedback is the pointwise regression that treats the target value of negative feedback as zero [13]. Here we opt for the pairwise way that directly encodes our ranking intuition, and leave the exploration of pointwise regression as future work.

Maximizing the objective function is equivalent to minimize the following loss function:

$$\mathcal{L} = \sum_{u \in \mathbb{U}} \sum_{i \in \mathbb{P}_u} \sum_{j \notin \mathbb{P}_u} \log \sigma(\hat{y}_{u,i} - \hat{y}_{u,j}) + \lambda_1 \left\| \mathbf{v}_u \right\|^2 + \lambda_2 \left\| \mathbf{q}_f \right\|^2, \tag{3}$$

where $\| \cdot \|$ denotes the L_2 norm for preventing model overfitting, and λ_1 and λ_2 are tunable hyper-parameters that control the extent of regularization.

As the number of training instances is very large (all user–item pairs) and there does not exist a closed form solution for model's parameters, learning is usually done by stochastic gradient descent (SGD). In each descent step, the localized optimization is performed on a tuple (u, i, j). The gradients with respect to each parameter are given as follows:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{v}_u} &= -\hat{e}_{u,i,j} \sum_{n=1}^{N} \left(\frac{1}{Z_{n,i}} \sum_{f \in F_{n,i}} \mathbf{q}_f - \frac{1}{Z_{n,j}} \sum_{f \in F_{n,j}} \mathbf{q}_f \right) + \lambda_1 \mathbf{v}_u, \\
\frac{\partial \mathcal{L}}{\partial \mathbf{q}^i_{n,f}} &= -\frac{1}{Z_{n,i}} \hat{e}_{u,i,j} \mathbf{v}_u + \lambda_2 \mathbf{q}^i_{n,f}, \\
\frac{\partial \mathcal{L}}{\partial \mathbf{q}^j_{n,f}} &= \frac{1}{Z_{n,j}} \hat{e}_{u,i,j} \mathbf{v}_u + \lambda_2 \mathbf{q}^j_{n,f},
\end{aligned} \tag{4}$$

where $\hat{e}_{u,i,j} = e^{-(\hat{y}_{u,i} - \hat{y}_{u,j})} / (1 + e^{-(y_{\hat{u},i} - y_{\hat{u},j})})$.

Then we iteratively loop over all the (u, i, j) tuples in the training set, and update the parameters to new values in the direction of negative gradient weighted by the learning rate until convergence. Learning rate is a key hyper-parameter for SGD that determines the speed of moving towards the optimal values — setting it too large we will skip the optimal solution, while too small a setting requires many iterations to converge. As such, we adopt *Bold Driver* [11], a technique that adjusts learning rate adaptively in each iteration. To be specific, it increases the learning rate by 5% if error rate is reduced since the last iteration; otherwise, resets the parameters to the values of the previous iteration and decreases the learning rate by 50%.

4.3.1 Time-aware Negative Sampling

In the literature of general recommender systems, uniform sampling is most widely used for sampling nagative instances due to its simplicity and acceptable performance [25]. Such sampling strategy assumes that all non-retweeted posts are equally weighted as negative instances (tweets disliked by user u). However, we believe this is invalid as: 1) many non-retweeted posts may simply be missed (never viewed) by user u, and 2) tweets may not be uniformly likely to have been seen by u. Previous works on tweet recommendation [33, 7, 15, 28] sampled negative instances from the tweets of the target user's followees only, assuming that the non-retweeted posts by the followees are more likely to be seen but disliked by the user. However, these previous efforts did not consider the effect of time, another important factor that determines whether the user may have seen the tweet.

To address this, we propose a time-aware negative sampling strategy. Our key assumption is that if a user has retweeted a post, she should also have read other tweets (of her followees) that were posted in close temporal proximity to the retweeted post. Such tweets are then more likely to be true negatives. Given a known image tweet interaction rt, we sample the non-retweeted image tweets (posted by her followees) in proportion to the time interval between the post and rt; *i.e.*, posts closer to rt have a higher chance of being selected. Our featured negative sampling method improves pairwise learning. We study its efficacy in Section 5.3.

Table 3: Image tweet training and test set demographics.

	Users	Retweets	All Tweets	Ratings
Training	926	174,765	1,316,645	1,592,837
Test		9,021	77,061	82,743

5. EXPERIMENT

We now evaluate CITING, our framework for context-aware image tweets modelling in the task of personalized image tweets recommendation. The goal of our experiment is to answer the following research questions:

RQ 1: How well do the four proposed contexts perform?

RQ 2: Do the filtered fusion improve model quality?

RQ 3: Can time-aware negative sampling strategy create better training set than uniform sampling?

RQ 4: Are visual objects sufficient to capture Twitter images' semantics?

Dataset. As there is no publicly available image tweet dataset, we constructed our own by gathering image tweets from Twitter in a user-centric manner. We first crawled one week of public timeline tweets (8–14 December 2014) which resulted in a set of 5,919,307 tweets, of which 17.2% contained images. From this collection, we randomly sampled 926 users who had at least 100 followees and 100–3000 followers, and posted at least 100 tweets. These requirements were used to select ordinary but active users, as has been done similarly by [28]. These 926 users are regarded as target users for our recommendation task.

We then crawled their latest tweets (up to 3,200 — limited by the Twitter API), their followee list and further crawled the image tweets published by their followees. In particular, given a user and her retweet rt, we sample 10 non-retweeted image tweets according to the time-aware negative sampling strategy described in Section 4.3.1. This process results in a dataset of 1,369,133 image tweets, summarized in Table 3. To simulate the real recommendation scenario, we adopt a time-based evaluation. For each user, we use her most recent 10 retweets as the test set, with the rest for training. Note that the user–tweet interaction is extremely sparse: each image tweet is retweeted by 1.22 users on average, and only 31% of the testing tweets have previously been observed in the training set. This validates the sparsity observations in previous works [7, 10].

Evaluation Metrics. The objective of tweet recommendation is to rank the candidate tweets such that the interesting tweets are placed at top for the target user. In our case, we mix the testing retweets (*i.e.,* ground-truth) and their negative samples as the candidate tweets for each user. To assess ranking quality, we adopt the average precision at

Table 4: Performance of each context source and its coverage (short for Cvr.). The best single context is the title of Google image search pages.

	P@1	P@3	P@5	MAP	Cvr.
P: Post	0.359	0.325	0.287	0.275	
P + Hashtag	0.360	0.324	0.293	0.277	14.3%
P + OCR text	0.366	0.332	0.301	0.283	26.4%
P + URL (title)	0.374	0.326	0.294	0.278	14.2%
P + URL (content)	0.381	0.330	0.300	0.279	13.2%
P + G (content)	0.369	0.319	0.289	0.275	57.2%
P + G (title)	**0.388**	**0.344**	**0.308**	**0.288**	76.0%
P + G (guess+NE)	0.381	0.330	0.296	0.280	58.1%

rank k (P@k) and Mean Average Precision (MAP) as evaluation metrics, which have been widely used for the tweet recommendation task [7, 10, 15]. Since users are usually most interested in only the few top recommendations, we report P@k at the top ranks (k=1, 3 and 5).

Parameter Settings. We tune two regularization parameters (λ_1 and λ_2) and the number of latent factors K. We first vary the regularizers until the results are generally stable, and then carefully tune K in a grid search manner (from 10 to 200). We report the performance at $\lambda_1 = 0.05$, $\lambda_1 = 0.01$ and $K = 160$, which shows good results. Similarly, we tune the parameters for other methods, and report their optimal results accordingly. For all the experiments, we set the initial learning rate as 0.01.

5.1 Utility of Proposed Contexts (RQ 1)

We first examine the efficacy of our proposed four strategies for context mining. To this end, we add the obtained text from each source to the post's original text separately, and assess performance using each combined text. For webpages, we separate the title and page content in evaluation, since we find some pages only have titles while lacking the main content and vice versa. Observing that some webpages can be very long and only the top few words (measured by *tf*idf*) are most relevant, we use the top 20 words[8] as the page content.

Table 4 shows the performance of each source with its coverage. In general, all context sources show a positive impact on the recommendation performance[9]. We find that the gains from the two external sources (external URL and Google Image Search) are more significant than the two internal sources (hashtag and OCR text). This validates the usefulness of external knowledge for interpreting images' semantics in social media. The largest improvement is obtained by integrating the titles of Google indexed pages, with a relative 8.1% and 4.7% improvement over using post's text only, in terms of P@1 and MAP, respectively. This improvement is partially due to the high coverage of Google Image Search over social media images. However, using the actual page content of the Google indexed pages neither improves over titles, nor betters the post's text — even degrading the performance for P@3 and MAP. Upon our deeper analysis, we find this might be caused by the noise introduced by Boilerpipe when extracting the main text from SNS pages and image aggregator sites. These sites make up a large portion in Google's indexed pages (84.3%) but their layouts significantly differ from news and blogs that Boilerpipe was trained on. As a result, Boilerpipe suffers from a high error rate. Thus in our subsequent experiments, we use all contextual text except the actual content of Google indexed pages.

[8]We experimented with a few settings (*e.g.*, 10, 20, 30) and found 20 works best.

[9] Although the P@3 slightly degrades for the source hashtag, other metrics still reveal it as a helpful feature.

Table 5: Context fusion performance comparison. '' denotes statistically significant difference vs. CITING with $p < 0.01$; '*' with $p < 0.05$.**

	P@1	P@3	P@5	MAP
(1): Random	0.114**	0.115	0.115	0.156**
(2): Length	0.176**	0.158	0.150	0.173**
(3): Profiling	0.336**	0.227	0.197	0.202**
(4): Post	0.359*	0.325	0.287	0.275**
(5): Non-filtered	0.413	0.352	0.319	0.296
(6): **CITING**	**0.419**	**0.355**	**0.319**	**0.298**

5.2 Effectiveness of Context Fusion (RQ 2)

We now evaluate the effectiveness of filtered fusion approach. For comparison purpose, we report the performance of our feature-aware MF model using all context without the filtered fusion (**Non-filtered**) and Post's text only (**Post**). The latter is equivalent to the results obtainable from two state-of-the-art models [7, 10] on text tweet recommendation, as the two are special cases of our model when only post's text is considered. To benchmark the performance, we also consider three baselines: 1) **Random**: ranking image tweets randomly; 2) **Length**: rank image tweets by the number of words in post's text, and the intuition is that longer tweets tend to be more informative and possibly to be more popular [32]; 3) **Profiling**: rank image tweets by the similarity of tweets' text and user's profile, which is constructed from the words of user's historical posts and retweets. To be specific, given a user u and an image tweet t, we compute the profile-based similarity score as follows:

$$S_{u,t} = (1 - w) \times cos(posts(u), t) + w \times cos(retweets(u), t),$$

where *cos* denotes the cosine similarity and w is a tunable parameter to balance the importance of the posting and retweeting history.

Table 5 shows the results. First, our proposed filtered fusion (CITING, R6) outperforms the three baselines (random, length, profiling) by a large margin. The filtered fusion method also significantly betters the strong baseline of using post's text by 0.06 (16.9% relative improvement) and 0.023 (8.3%) in terms of P@1 and MAP, respectively. When adopting non-filtered fusion approach, the performance slightly drops, *e.g.*, the P@1 drops from 0.419 to 0.413. Although not statistically significant, it indicates that our heuristic filtered fusion approach achieves comparable results while saving acquisition costs of the contextual text by 18.0%. These experimental results evidence the effectiveness of our fusion approach and the feature-aware MF model.

5.3 Importance of Negative Sampling (RQ 3)

We now assess the effect of the negative sampling strategy. We compare with the uniform sampling strategy, which is a commonly used strategy by previous works in tweet recommendation [33, 7, 28]. To this end, we constructed a new dataset by uniformly sampling negative image tweets from our training set and pair with the positive image tweets. We then trained our feature-aware MF on this new dataset, us-

Table 6: Performance using visual objects.

	P@1	P@3	P@5	MAP
(1): CITING	0.419	0.355	0.319	0.298
(2): Visual objects (V)	0.221	0.205	0.192	0.211
(3): Post's text + V	0.379	0.325	0.293	0.280
(4): CITING + V	0.425	0.350	0.313	0.298

ing our proposed filtered contexts, and evaluated the method in the same way. Experimental result shows the time-aware sampling strategy significantly betters the random sampling by 0.017 (4.2% relative improvement) and 0.006 (2.1%), for P@1 and MAP, respectively. Both one-tailed paired t-test for P@1 and MAP show $p < 0.05$. This validates our time-aware negative sampling strategy is effective in constructing a higher quality training set, aiding better user interest modelling.

5.4 Insufficiency of Visual Objects (RQ 4)

We now validate our claim at the outset that annotating visual objects without context does not fare well for social media interpretation. First, we applied *GoogLeNet* [27], the winning system in ILSVRC 2014, to classify the visual objects for our Twitter images. GoogLeNet is trained on 1.2 million Flickr images with 1000 object categories, where each category corresponds to a node in ImageNet/WordNet. The pre-trained model is provided by Caffe [17]. We take the top five labels as the description for each image and conduct the same experiment. We see that prediction using just visual objects does perform worse (P@1= 0.221, MAP= 0.211; Table 6, R2), due to the largely literal image descriptions. Our CITING context significantly outperforms visual objects by 89.2% of relative improvement and 40.9% in terms of P@1 and MAP, respectively. This shows the contextual text does capture image tweets' semantics much better.

For comprehensiveness, we further experiment with the combination of text and visual objects (*i.e.*, model the two as two types of features), to see whether the incorporation of visual cues could further boost the recommendation performance. As shown in Table 6 (R3), the integration of visual objects with post's text slightly improves over post's text 5.6% (relative improvement) and 1.8%, for P@1 and MAP, respectively, while our CITING context still significantly betters such combination by relatively 10.5% and 6.2%. This further validates our contextual text is able to capture semantics of image tweets better. Unlike the previous combination, the incorporation of visual objects does not lead to a stable improvement for contextual text: P@1 is slightly improved by 0.006 (1.4%), and MAP remains the same, while the other two metrics drop. This suggests the descriptors brought about by using visual objects is limited in modelling usefulness, and such visual cues might have already been largely captured by our contextual text (*e.g.* some best guess descriptions from Google Image Search describe visual objects).

5.5 Case Studies

It is also instructive to examine individual users and actual posts to better understand the effectiveness of our proposed filtered contextual text. To this end, we examine a few users whose recommendations have a large performance gain when using CITING. In Figure 9, we show such a typical user (refer as *User 1*) and four of her retweets in test set that are

Figure 9: Four image tweets from *User 1*'s retweets in test set benefit from our contextual text.

Figure 10: Five image tweets from *User 2*'s retweets in test set benefit from our contextual text.

enriched by our contextual text. As a consequence, the average recommendation precision of our approach (0.592) significantly outperforms the approach of using visual objects (0.226) and using post's text (0.443). In an even more successful case, 9 out of 10 retweets (we show 5 in Figure 10) for *User 2* obtained contextual text from our approach. The average precision is boosted from 0.423 (using visual objects) and 0.319 (using post's text) to 0.728 (our approach).

Taking a closer look at the these image tweets, we find a few of them trigger multiple context mining channels. Some have both embedded URL and overlaid text in image (see Figure 10: the top leftmost and the bottom rightmost). A further investigation shows the external web pages redirected by embedded URLs contain richer and more relevant information than the overlaid text. This validates our text fusion strategy which assigns a higher priority to text from embedded URL than OCR. Another image tweet (Figure 9, top leftmost) has both overlaid text and search result from Google Image. However, the search result only indicates the image is a quote, but does not reveal its deep semantics as OCR text does. In this case, OCR text is more reliable than search result, validating our text fusion strategy 2.

6. CONCLUSION

Compared to the traditional vision research on stock photo images, we have shown that social media images are more semantic but diverse, which need to be understood within their context of mention. To complement the visual features, we propose a CITING framework that mines both the intrinsic and extrinsic contexts of image tweets. To demonstrate the effectiveness, we focus on the task of personalized image tweet recommendation, developing a feature-aware recommendation system that encodes the contexts as part of user interest modelling. Extensive experiments verify the effectiveness of our proposed CITING method in context mining, significantly boosting recommendation performance.

We have done an analysis of the coverage and efficacy of acquiring the textual context of social media images, but there is still much that can be improved here. To spur additional research on social media images, we have released the code of our feature-aware MF model, our large image tweets dataset, as well as our annotated corpus of 500 sample images with their manually-recognized text[10]. In particular, future work can adapt OCR to better acquire text within the images, as current OCR fares poorly on meme-style images and graphics. Additionally, we plan to examine whether other types of features (*e.g.*, geo-location or publisher) would result in even better user modelling.

Acknowledgments

This research is supported by the Singapore National Research Foundation under its International Research Centre. We also would like to thank Yongfeng Zhang and Hanwang Zhang for their help and discussions.

7. REFERENCES

[1] O. Alonso, C. C. Marshall, and M. Najork. Are Some Tweets More Interesting Than Others? #HardQuestion. In *Proc. of HCIR*, 2013.

[2] J. Bian, Y. Yang, and T.-S. Chua. Multimedia Summarization for Trending Topics in Microblogs. In *Proc. of CIKM*, 2013.

[3] J. Bian, Y. Yang, and T.-S. Chua. Predicting Trending Messages and Diffusion Participants in Microblogging Network. In *Proc. of SIGIR*, 2014.

[4] H. Cai, Y. Yang, X. Li, and Z. Huang. What Are Popular: Exploring Twitter Features for Event Detection, Tracking and Visualization. In *Proc. of MM*, 2015.

[5] V. Campos, A. Salvador, X. Giro-i Nieto, and B. Jou. Diving Deep into Sentiment: Understanding Fine-tuned CNNs for Visual Sentiment Prediction. In *Proc. of ASM*, 2015.

[6] E. F. Can, H. Oktay, and R. Manmatha. Predicting Retweet Count Using Visual Cues. In *Proc. of CIKM*, 2013.

[7] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu. Collaborative Personalized Tweet Recommendation. In *Proc. of SIGIR*, 2012.

[8] T. Chen, D. Lu, M.-Y. Kan, and P. Cui. Understanding and Classifying Image Tweets. In *Proc. of MM*, 2013.

[9] T. Chen, H. M. SalahEldeen, X. He, M.-Y. Kan, and D. Lu. VELDA: Relating an Image Tweet's Text and Images. In *Proc. of AAAI*, 2015.

[10] W. Feng and J. Wang. Retweet or Not?: Personalized Tweet Re-ranking. In *Proc. of WSDM*, 2013.

[11] R. Gemulla, E. Nijkamp, P. J. Haas, and Y. Sismanis. Large-scale Matrix Factorization with Distributed Stochastic Gradient Descent. In *Proc. of SIGKDD*, 2011.

[10]http://wing.comp.nus.edu.sg/downloads/image-tweet-ocr-rec

[12] X. He, T. Chen, M.-Y. Kan, and X. Chen. TriRank: Review-aware Explainable Recommendation by Modeling Aspects. In *Proc. of CIKM*, 2015.

[13] X. He, H. Zhang, and M.-Y. Kan. Fast Matrix Factorization for Online Recommendation with Implicit Feedback. In *Proc. of SIGIR*, 2016.

[14] L. Hong, O. Dan, and B. D. Davison. Predicting Popular Messages in Twitter. In *Proc. of WWW*, 2011.

[15] L. Hong, A. S. Doumith, and B. D. Davison. Co-factorization Machines: Modeling User Interests and Predicting Individual Decisions in Twitter. In *Proc. of WSDM*, 2013.

[16] K. Ishiguro, A. Kimura, and K. Takeuchi. Towards Automatic Image Understanding and Mining via Social Curation. In *Proc. of ICDM*, 2012.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proc. of MM*, 2014.

[18] C. Kohlschütter, P. Fankhauser, and W. Nejdl. Boilerplate Detection Using Shallow Text Features. In *Proc. of WSDM*, 2010.

[19] Y. Koren, R. Bell, and C. Volinsky. Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8):30–37, 2009.

[20] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *Proc. of WWW*, 2010.

[21] D. Laniado and P. Mika. Making Sense of Twitter. In *Proc. of ISWC*, 2010.

[22] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter. In *Proc. of WebSci*, 2011.

[23] S. Petrovic, M. Osborne, and V. Lavrenko. RT to Win! Predicting Message Propagation in Twitter. In *Proc. of ICWSM*, 2011.

[24] S. Rendle. Factorization Machines. In *Proc. of ICDM*, 2010.

[25] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proc. of UAI*, 2009.

[26] M. Schinas, S. Papadopoulos, Y. Kompatsiaris, and P. A. Mitkas. Visual Event Summarization on Social Media Using Topic Modelling and Graph-based Ranking Algorithms. In *Proc. of ICMR*, 2015.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. In *Proc. of CVPR*, 2015.

[28] I. Uysal and W. B. Croft. User Oriented Tweet Ranking: A Filtering Approach to Microblogs. In *Proc. of CIKM*, 2011.

[29] A. Wang, T. Chen, and M.-Y. Kan. Re-tweeting from a Linguistic Perspective. In *Proc. of LSM*, 2012.

[30] Z. Wang, P. Cui, L. Xie, H. Chen, W. Zhu, and S. Yang. Analyzing Social Media via Event Facets. In *Proc. of MM*, 2012.

[31] Z. Wang, P. Cui, L. Xie, W. Zhu, Y. Rui, and S. Yang. Bilateral Correspondence Model for Words-and-Pictures Association in Multimedia-Rich Microblogs. *ACM TOMM*, 10(4):34:1–34:21, 2014.

[32] R. Yan, M. Lapata, and X. Li. Tweet Recommendation with Graph Co-ranking. In *Proc. of ACL*, 2012.

[33] T. R. Zaman, R. Herbrich, J. V. Gael, and D. Stern. Predicting Information Spreading in Twitter. In *Computational Social Science and the Wisdom of Crowds Workshop*, 2010.

[34] H. Zhang, G. Kim, and E. P. Xing. Dynamic Topic Modeling for Monitoring Market Competition from Online Text and Image Data. In *Proc. of KDD*, 2015.

[35] H. Zhang, X. Shang, H. Luan, M. Wang, and T.-S. Chua. Learning from Collective Intelligence: Feature Learning Using Social Images and Tags. *ACM TOMM*, 2016.