# Multimodal Context-Aware Recommender for Post Popularity Prediction in Social Media

Masoud Mazloom
Informatics Institute
University of Amsterdam
m.mazloom@uva.nl

Bouke Hendriks
Informatics Institute
University of Amsterdam
boukehendriks@gmail.com

Marcel Worring
Informatics Institute
University of Amsterdam
m.worring@uva.nl

## ABSTRACT

Millions of multimodal posts are uploaded, shared, viewed and liked every day in different social networks, where users express their opinions about different items such as products and places. While, some user posts become popular, others are ignored. Even different posts related to the same items shared by different users receive a different number of likes and views. Existing research on popularity prediction aggregate all user posts related to different items without considering the preferences of individual user for the items in training a popularity model. This often results in limited success. We hypothesize that popularity of posts differs from one user to the other user, one item to the other items, and posts related to similar users or similar items may receive the same number of likes. In this paper, we present an approach for predicting the popularity of user posts by considering preferences of individual users to the items. We factorize the popularity of posts to the user-item-context and propose a multimodal context-aware recommender. Using our proposal we have the ability of predicting the popularity of posts related to different items which are shared by a specific user. Moreover we are able to predict the popularity of posts shared with different users for a specific item. We evaluate our approach on an Instagram user posts dataset with over 600K posts in total related to different touristic places in The Netherlands, as items, for the task of popularity prediction.

**ACM Reference format:**
Masoud Mazloom, Bouke Hendriks, and Marcel Worring. 2017. Multimodal Context-Aware Recommender for Post Popularity Prediction in Social Media. In *Proceedings of ThematicWorkshops'17, October 23–27, 2017, Mountain View, CA, USA.*, , 9 pages.
DOI: https://10.1145/3126686.3126731

## 1 INTRODUCTION

In recent years predicting the popularity of user generated posts in social networks has attracted attention because of its widespread applications such as content recommendation [4], advertisement [14], information retrieval [29], and online marketing [26]. In this paper, we focus on the problem of predicting the popularity of user posts related to different items which are shared in social networks.

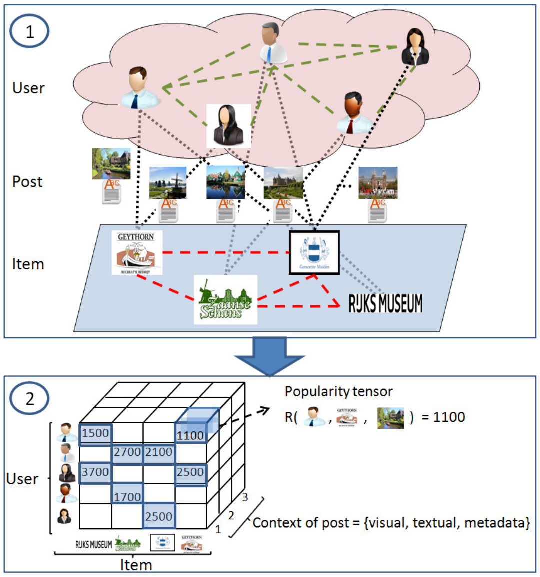

**Figure 1: We formulate the popularity prediction of user posts related to specific items as a multimodal context-aware recommender. Step 1 shows that users interact with items, here touristic places in The Netherlands, by sharing multimodal posts in social media. Step 2 shows the popularity tensor by considering users, items, and contexts of posts which we use for predicting the popularity of a post.**

We define items here as the topics of posts. It can be a product of a specific brand, *e.g.* Nike shoes, a brand itself, *e.g.* Nike, or a touristic place, like Amsterdam. While, in particular, some user posts intrinsically become more popular, others are ignored. Even posts related to different items shared by the same user receive different number of likes and views. The question arises what is it that makes a post shared by a specific user related to a specific item become more popular? Can we predict the number of likes a post will receive for a specific user and specific item? The information to answer this question is explicitly captured in parameters such as the number of likes or views, and implicitly hidden in the text and visual content of the posts which have been shown to be powerful elements to draw attention in social media [18]. In order to benefit from the information in user posts for investigating post popularity, we need to leverage both the explicit information in a post as well as the visual and textual context.

Recently several works propose to predict the popularity of user generated contents in social media [1, 6, 9, 12, 18, 19]. In [1, 9], the authors consider the problem of popularity prediction on textual content specifically predicting the popularity of tweets using textual

*ThematicWorkshops'17, October 23–27, 2017, Mountain View, CA, USA.*
© 2017 ACM. ISBN 978-1-4503-5416-5/17/10…$15.00
DOI: https://10.1145/3126686.3126731

features in Twitter. The authors in [6, 12, 18, 19] investigate the effect of different visual features for predicting the popularity of an image. All these works learn a popularity model directly from the visual and textual features of a post without considering the interaction between the user generated post and the item the post is about. While some works [21, 27, 28] incorporate user information, we believe we are the first to model popularity from the user-item-context of posts.

In this paper we propose a novel framework based on multimodal context-aware matrix factorization, where the popularity is estimated through a user-item-context popularity tensor. Our proposal relies on the information of how popular a *post* is with its visual and textual *context* which is shared by a *user* related to an *item*. We use a wide variety of context domains including information related to user, item, visual, and textual information related to the content of a post shared by a user. Our context-aware predictor is based on *Factorization Machine* (FM) [22] which allows fast prediction and learning with context-aware data. We extend FM by using visual and textual contents as information. By formulating popularity prediction as a context-aware recommender we are keeping the dependency of users, items, and as well as user-item interaction in the prediction of popularity of a post. By this, we have the ability of predicting the popularity for a specific-user and -item. For a specific user who has a big collection of images it is important to select a small set of images for sharing in social networks which are most likely to receive a high number of likes. For a specific item, such as a tourist organization or brand companies, it is also important to select and share images of their products as advertisement in their fan page in social networks which their costumers would like most. Figure 1 visualizes how we can map the popularity of a post into a user-item-context tensor. We make the following main contributions in this paper:

- We address the problem of specific-item popularity prediction of user posts.
- We formulate the popularity prediction in terms of a user-item-context popularity tensor and propose a new representation of a post.
- We propose a multimodal context-aware recommender for predicting popularity.
- Our experiments show empirically that our proposal improves the popularity predictive accuracy.
- We introduce a new Instagram based dataset for popularity prediction.

We organize the remainder of this paper as follows. We begin by mentioning related work in Section 2. Section 3 describes our problem formulation and we define our proposal for predicting the popularity of a post. We introduce the experimental setup on our dataset in Section 4. Results are presented in Section 5. Finally, Section 6 concludes with a summary of our findings and a discussion of several possible directions for future work.

## 2 RELATED WORK

The earlier work on popularity prediction of user generated posts in social media has focused on the textual content of posts such as tweets on Twitter [1, 9] or user comments on web items [8]. Hong *et al.* in [9] predict the popularity of tweets and number of retweets, in Twitter by textual features, *e.g.* sentiment, extracted from tweets. They emphasize the effect of combining textual features with contextual features of the user. In [1] Bae *et al.* show the effectiveness of textual sentiment analysis in predicting the popularity of tweets. He *et al.* in [8] analyze user comments for predicting the popularity of web 2.0 items. By modeling user comments as a time-aware bipartite graph, they propose a ranking algorithm that takes into account temporal information to predict the future popularity of items. All these works [1, 8, 9] predict the popularity of users posts using textual features, the effect of visual content which is rich in information is not addressed.

Recently, a significant effort has been spent on such use of visual content as an additional channel in prediction. In [3, 6, 12, 18, 19], the authors focus on visual and textual content for predicting the popularity of a post and measuring the impact of different visual and textual features on popularity, as well as their combination. Khosla *et al.* in [12], report the results of image popularity prediction using different visual features and contextual features. They consider the problem of popularity prediction as a learning to rank problem. McParlane *et al.* in [19], report the effect of social factors of each post such as how many followers a user has, the number of tags attached to the photo, and the length of the title. Moreover, they report on the result of popularity prediction using content factors such as the number of faces in the images, analysis of the scene, and color features. They investigate the problem of popularity prediction of a post as a binary classification. Cappallo *et al.* in [3] learn a ranker by considering popular and unpopular latent factors. The authors in [6, 18] investigate the effect of sentiment analysis on the popularity of a post. Gelli *et al.* in [6] investigate the effect of visual sentiment analysis on images as well as the contextual features used in [19]. They report the potential of predicting the popularity of images using visual sentiment scores as a feature, which was first introduced in [2] to detect sentiment in an image. Mazloom *et al.* in [18] initiate the problem of predicting popularity of brand-related user posts automatically in the business and marketing community. They further propose usage of an ensemble of cues, extracted from the visual and textual channel of posts, which are important in analyzing brand popularity. All these works [6, 12, 18, 19] emphasize that visual and textual features are complementary in predicting the popularity of user generated posts. However, all these works [3, 6, 12, 18, 19] model the popularity of posts of different users related to different items by jointly considering all posts without the dependency between users and items inside the model. By ignoring information related to user-item interaction, these works have the limitation that they are not capable of jointly predicting popularity for a specific user and a specific item which often results in limited success. Different from all these works, we propose to investigate the popularity prediction of posts by factoring popularity into the user-item-context (visual and textual content of a post) which is linked to the popularity for a specific user and a specific item.

In [21, 27, 28] the authors use the user-item interaction information derived from user-item sharing behaviours on social media to predict the popularity of an image. In [21], Niu *et al.* consider the user-item interaction as a weighted bipartite graph to model the popularity of a post. They use social factors, such as the number of click-throughs and number of comments for presenting the post

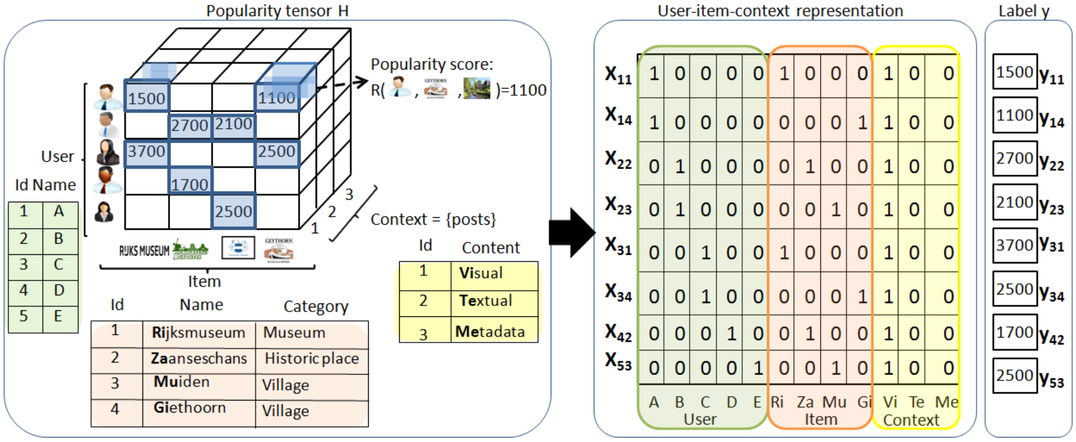

**Figure 2: Popularity tensor data (left side) is transformed into a user-item-context representation of posts with their labels (right side). In the new representation, the first five values indicate the user, the next four the items, and the last three indicate visual, textual, or metadata features as the context of a post in a multimodal context-aware recommender. The element $x_{11}$ states that the post shared by user $A$ related to item *Rijksmuseum* by considering the visual content of the post got 1500 likes.**

without considering the visual and textual content of posts. Wu *et al.* propose to predict the popularity of a post in Flickr by unfolding the user-item contextual dynamics and incorporating temporal context into the prediction. They consider visual features and social factors as contextual information. However their method has some limitations: firstly textual contents and sentiment analysis of posts, both effective factors for post popularity prediction[7, 18], are ignored. Secondly, they consider the images of each user as items, which leads to a computationally intensive solution. Finally, they rely on images and not specific items such as brand products or touristic places as items. Consequently, they are limited to predicting the popularity of user specific posts, not item specific posts. In this paper, we use visual, textual content and sentiment of user posts related to specific items as contextual information and propose to model the popularity of a post by a multimodal context-aware recommender model.

# 3 POPULARITY FROM USER-ITEM-CONTEXT

We view the popularity of a multimodal post as a context-aware signal, which is highly correlated to who creates the post, the user, the user interest, the item, and the visual and textual context of a post. In this paper, we aim to predict the popularity of a post related to a specific-user and -item in social media. For this purpose, we propose a multimodal context-aware recommender which takes into account the interaction of the user, the item and the context of the post for modeling its popularity. Before introducing the proposed framework for popularity prediction, the notation and key concepts will be formally introduced.

## 3.1 Problem Formalization

Given a specific-user and -item, popularity prediction is the task of computing a score that predicts how popular the post will be in comparison to the other posts of the user and the item. We use $p_{sj}$ to indicate the post generated by the $s^{th}$ user $u_s$, related to the $j^{th}$ item $i_j$. We aim to construct a real-valued function, $\hat{y}_{sj} = R(u_s, i_j, p_{sj})$, which estimates the popularity score for $p_{sj}$. By sorting all posts in a test set according to $R()$ in descending order, a list of most popular posts will be obtained for both specific-user and -item.

Let $U = \{u_1, u_2, \ldots, u_m\}$ be a set of $m$ users, $I = \{i_1, i_2, ..., i_r\}$ be a set of $r$ items, and $P = \{p_{11}, p_{12}, \ldots, p_{mr}\}$ a set of $d$ posts, with $d \ll m * r$, which users generated and shared in social media related to items. Note that each post $p_{sj} = \langle p^v_{sj}, p^t_{sj} \rangle$ consists of two modalities, visual $p^v_{sj}$, and textual $p^t_{sj}$ content. Suppose $y_{sj}$ is the ground truth popularity score of post $p_{sj}$, which is the number of likes this post has received in social media. Then we define $H = \{(p_{11}, y_{11}), (p_{12}, y_{12}), ..., (p_{mr}, y_{mr})\}$ as a dataset of $d$ labeled posts. We use dataset $H$ for learning function $R()$, as a popularity predictor.

The difficulty in constructing $R()$ for estimating the popularity score largely depends on the representation of the triplet *(user, item, post)*. We hypothesize that what makes a post popular on the web depends on the context of the post, the user which shared the post, and the item the post is about. The key idea of our proposal is to learn function $R()$ based on the user-item-context representation of a post.

Next, we show in section 3.2 how to construct a representation based on the triplet of user, item, and context of post. For the ease of reference, Table 1 lists the main notation used throughout this work.

## 3.2 User-item-context Representation

A mathematical representation of the popularity dataset $H$ is defined with a three-dimensional tensor, which we call popularity tensor, with the user dimension, item dimension, and context dimension (See Figure 2). Then the popularity prediction of a post can be formulated as predicting unobserved popularity entries in the popularity tensor $H$ based on other observed data in a context-aware recommender system manner.

We propose a *user-item-context representation* of each post as the input for our multimodal context-aware recommender proposal. The context-aware recommender assumes the contextual information is known and defined by a set of contextual features which

**Table 1: Main notations used in this work**

| Notation | Definition |
|---|---|
| $U$ | a set of $m$ users |
| $I$ | a set of $r$ items |
| $p_{sj}$ | a post shared by user $s$ related to item $j$ |
| $x_{sj}$ | *user-item-context* representation of $p_{sj}$ |
| $y_{sj}$ | number of likes which $p_{sj}$ received |
| $P$ | a set of $d$ posts users shared about items |
| $H$ | a labeled dataset of $\{(x_{sj}, y_{sj})\}$ |
| $R()$ | a function for computing the popularity |

affect the popularity score. In other words, the popularity is modeled in the *user-item-context representation* approach to context-aware recommendation as the function of users, items, and the contextual features of post as:

$$R : D_1 \times D_2 \times D_3 \longrightarrow Popularity, \tag{1}$$

where $D_1$ and $D_2$ are the domains of users $U$ and items $I$ respectively, *Popularity* is the domain of popularity, and $D_3$ specifies the contextual information of post. Each dimension $D_i$ is a subset of some attributes. In this paper we defined $D_1 \subseteq UserId \times UserName$. Similarly, $D_2 \subseteq ItemId \times ItemName \times ItemType$. Finally, the *Context* dimension can be defined as $D_3 \subseteq Contextualfeatures$ of posts, $D_3 \subseteq D_v \cup D_t$, where $D_v$ and $D_t$ are dimensions of visual and textual features of posts respectively. Figure 2 (right side) show our *User-item-context* representation of each post and the popularity labels. We represent the popularity tensor $H$ as $H = \{(x_{11}, y_{11}), (x_{12}, y_{12}), ..., (x_{mr}, y_{mr})\}$ where $x_{sj} \subseteq \mathbb{R}^{D_1 \times D_2 \times D_3}$ is a *User-item-context* representation of $p_{sj}$.

After representing posts, the next step is to learn the popularity function $R()$ to estimate the unknown popularity of posts.

## 3.3 Multimodal Context-aware FMs for Popularity Prediction

Let $x \in \mathbb{R}^n, n = D_1 \times D_2 \times D_3$ denote a *User-item-context* representation of a post in test time. The goal of this section is to learn the parameter vector $\Theta$ of function $R()$ over training set $H$ to estimate the popularity of $x$, $\hat{y}(x)$.

Inspired by the success of Factorization Machine (FM) [22] in a context-aware recommender system [23], we propose to use FM in our multimodal context-aware recommender which models all interactions between pairs of variables, $var = \{U, I, VC, TC\}, VC \subseteq D_v, TC \subseteq D_t$, with the popularity including nested ones, by factorized interaction parameters:

$$\hat{y}(x) = R(x) = w_0 + \sum_{a=1}^{z} w_a x_a + \sum_{a=1}^{z} \sum_{b=a+1}^{z} \hat{w}_{a,b} x_a x_b \tag{2}$$

where $z$ is the number of variables, $z = |var|$, $\hat{w}_{a,b}$ are the factorized interaction parameters between variable pairs:

$$\hat{w}_{a,b} = \langle v_a, v_b \rangle = \sum_{f=1}^{k} v_{a,f} . v_{b,f} \tag{3}$$

where $k$ is the dimensionality of the factorization and the model parameter vector $\Theta$ which has to be estimated is composed of:

$$w_0 \in \mathbb{R}, \quad \mathbf{w} \in \mathbb{R}^z, \quad \mathbf{V} \in \mathbb{R}^{z \times k} \tag{4}$$

**Algorithm 1** Learning the model parameters of our proposal, $w_0$, $\mathbf{w}$ and $\mathbf{V}$ using ALS algorithm.

```
procedure LearningParameters(training set H)
    ▹ Initialize the model parameters
    $w_0 \leftarrow 0, \mathbf{w} \leftarrow (0, \cdots, 0), \mathbf{V} \sim \mathcal{N}(0, \sigma)$
    ▹ Precompute error $e$ and factorized parameter $q$
    for $(x, y) \in H$ do
        $e(x, y|\Theta) \leftarrow \hat{y}(x) - y$
        for $f \in \{1, \cdots, k\}$ do
            $q(x_{ij}, f|\Theta) \leftarrow \sum_{a=1}^{z} v_{a,f} x_a$
        end for
    end for
    ▹ Main optimization loop
    repeat
        ▹ global bias
        $w_0^* \leftarrow -\frac{\sum_{(x,y) \in H} (e(x,y|\Theta) - w_0)}{|H| + \lambda_{(w_0)}}$
        $e(x, y|\Theta^*) \leftarrow e(x, y|\Theta) + (w_0^* - w_0)$
        $w_0 \leftarrow w_0^*$
        ▹ Interaction of each variable
        for $l \in \{1, \cdots, z\}$ do
            $w_l^* \leftarrow -\frac{\sum_{(x,y) \in H} (e(x,y|\Theta) - w_l x_l) x_l}{\sum_{(x,y) \in H} x_l^2 + \lambda_{(w_l)}}$
            $e(x, y|\Theta^*) \leftarrow e(x, y|\Theta) + (w_l^* - w_l) x_l$
            $w_l \leftarrow w_l^*$
        end for
        ▹ Interaction of pair of variables
        for $f \in \{1, \cdots, k\}$ do
            for $l \in \{1, \cdots, z\}$ do
                $v_{l,f}^* \leftarrow -\frac{\sum_{(x,y) \in H} (e(x,y|\Theta) - v_{l,f} h_{(v_{l,f})}(x))}{\sum_{(x,y) \in H} h_{(v_{l,f})}^2(x) + \lambda_{(v_{l,f})}}$
                $e(x, y|\Theta^*) \leftarrow e(x, y|\Theta) + (v_{l,f}^* - v_{l,f}) x_l$
                $q(x, f|\Theta^*) \leftarrow q(x, f|\Theta) + (v_{l,f}^* - v_{l,f}) x_l$
                $v_{l,f} \leftarrow v_{l,f}^*$
            end for
        end for
    until stopping criterion is met
    return $w_0, \mathbf{w}, \mathbf{V}$
end procedure
```

where $w_0$ is the global bias, $w_a$ models the interaction of the $a^{th}$ variable to the popularity and $\hat{w}_{a,b}$ models the factorized interaction of a pair of variables with the popularity.

Formula 2 can be computed very efficiently as it is equivalent to:

$$\hat{y}(x) = w_0 + \sum_{a=1}^{z} w_a x_a + \frac{1}{2} \sum_{f=1}^{k} \left( \left( \sum_{a=1}^{z} v_{a,f} x_a \right)^2 - \sum_{a=1}^{z} v_{a,f}^2 x_a^2 \right) \tag{5}$$

We use the following regularized least square criterion to prevent overfitting and optimized the parameter vector $\Theta$:

$$\sum_{(x_{ij}, y_{ij}) \in H} (\hat{y}(x_{ij}) - y_{ij})^2 + \sum_{\theta \in \Theta} \lambda_{(\theta)} \Theta^2 \tag{6}$$

where $\lambda_{(\theta)}$ is a regularization hyper-parameters for the model parameter $\theta$.

**Learning Algorithm** The model equation in formula (6) can be calculated in linear time. We use the alternating least square (ALS) learning algorithm in [22] which finds the optimal value for a model parameter given the remaining ones very quickly. By ALS, a joint optimum of all parameter elements Θ can be found iteratively by calculating the optimum of each model parameter one after another and repeating this several times. Our learning algorithm for obtaining the optimum value of Θ is summarized in Algorithm 1. First the model parameters are initialized, where $w_0$ and $w_l$, the interaction of the $l^{th}$ variable to the popularity can be initialized with 0 and the factorization parameters with small 0-centered random values. In the main loop the parameters are optimized one after the other. This optimization main loop is repeated several times to converge to the joint optimum of all model parameters. In Algorithm 1, function $h_{(\theta)}(x) = \frac{\partial}{\partial\theta}\hat{y}(x|\theta)$ depends on the variable $\theta$, but is not dependent on the value of $\theta$. $e \in \mathbb{R}^{|H|}$ is a vector of errors over all training examples and it is precomputed. After storing the error terms, the computation complexity only depends on the complexity of the $h_{(\theta)}$ functions. For the factorized parameters, computing $h$ contains a loop over all variables. For this reason we define variable $q$ in our algorithm which can be precomputed for each training case and factor in a matrix $Q \in \mathbb{R}^{|H| \times k}$. With a precomputation of the $q$-terms, the $h$ function can be computed in constant time.

# 4 EXPERIMENTAL SETUP

We evaluate the effectiveness of our multimodal context-aware recommendation system for predicting the popularity of a post by performing a series of experiments on a dataset crawled from Instagram.

## 4.1 Dataset

Since there is no dataset in popularity prediction which considers the user-item-post interaction, we construct a dataset by crawling Instagram. Instagram is chosen as a data collection platform, as it has a strong focus on self-expression and also offers a vast amount of publicly available multimodal data. We target popularity prediction of posts related to touristic places as items, in particular from one particular city namely the Amsterdam Metropolitan Area. Following, we describe our method for constructing our dataset [1].

The first step involves defining a list of *items* found within the Amsterdam Metropolitan Area by Amsterdam Open Data [15] such as "*anne frank huis*"). We utilise the Instagram API [10] to crawl the Instagram posts related to *items*. For every $item$ we collect a set of posts. The returned data is denoted as $D_{initial}$, which holds all posts related to every *item*. The oldest post in our dataset is dated 28-Oct-2010 and the most recent post is dated 07-Apr-2016.

In our experiments, we limit our dataset to items that have a minimum of 500 posts, and keep those users which shared posts related to at least 10 different items. We also consider only those posts which have both visual and textual information. The final dataset $D_{clean}$ which we used in our experiment contains 599,756 multimodal posts related to 152 items which are shared by 14,347 unique users. The statistics of our dataset are shown in Table 2. We split the dataset randomly into train, 60%, and test set, 40%.

**Table 2: The statistic of our Instagram dataset. The final dataset is marked with $^*$ after cleaning up procedure.**

| Term | Size |
|---|---:|
| $items$ | 472 |
| unique users | 426238 |
| $D_{initial}$ | 3129709 |
| $r$: items with ≥ 500 posts | 152$^*$ |
| $m$: users shared posts related to ≥ 10 items | 14,347$^*$ |
| $D_{clean}$ | 599,756$^*$ |

In order to explore different data distributions that occur in various applications and social networks, we evaluate our proposal in two different settings namely *user-specific*, and *item-specific*.

**user-specific**: For this setting, we randomly select 1000 users from the test set who shared posts related to at least 30 different items. We train a model on the train set and report the popularity score for each of the selected users on the test set and average the result.

**item-specific**: In this setting, we randomly select 50 items from the test set of users which shared more than 30 posts related to them. We used the trained model on the train set and report the average result of popularity score over all 50 items.

## 4.2 Implementation details

**Contextual features** As indicated before, we consider the visual and textual information of a post as contextual information. To extract and represent the contextual information from both the visual and textual content of a post, we use the following state of the art features, all in line with [18]:

*Textual features*:

- *W2V* A trained deep neural network proposed in [20], which computes a 300 dimensional vector for each tag using average pooling of the W2V representation of all tags to represent the post.
- *Textual Sentiment* Represented by making use of SentiStrength [25] which generates a positive and negative sentiment score per tag.

*Visual features*:

- *CNN-Pool5* We use the 1024-dimensional features from pooling the last fully connected layer of the Deep Net in [24], trained on ImageNet [5].
- *Concepts* We represent the image of each post by the 15,293-dimensional output of the softmax layer of the Deep Net [24] trained to identify 15,293 ImageNet [5] concepts as it has been shown to be an effective representation in [16].
- *Visual Sentiment* A representation of a post based on the Visual Sentiment Ontology [2] to detect sentiment in an image. This representation consists of a 1,200 dimensional vector with probabilities of Adjective Noun Pairs (ANP) being present in the image.

**Learning algorithm parameters** In formula (6) we set the regularization parameters: $\lambda_{w_0} = 0$ as there is no need to regularize the global bias, the same $\lambda_w = 0.01$ for all parameters $w_l \in \mathbf{w}$ and the same $\lambda_v = 0.1$ for all parameters $v_{a,f} \in \mathbf{V}$. We consider 10% of the training set as validation set for tuning these parameters for

[1] http://isis-data.science.uva.nl/Masoud/MM17Data.

**Table 3: Post popularity prediction on user-specific setting.**

| | **Textual Features** | | **Visual Features** | | | **Multimodal** |
|---|---|---|---|---|---|---|
| | W2V | Sentiment | CNN-Pool5 | Concept | Sentiment | **Late Fusion** |
| PKNN | 0.177 | 0.084 | 0.195 | 0.114 | 0.164 | 0.206 |
| PSVR | 0.462 | 0.281 | 0.489 | 0.394 | 0.429 | 0.511 |
| PRSVM | 0.431 | 0.227 | 0.443 | 0.402 | 0.410 | 0.475 |
| PBG | 0.455 | 0.241 | 0.473 | 0.369 | 0.414 | 0.495 |
| PCFM(ours) | **0.501** | **0.311** | **0.521** | **0.412** | **0.470** | **0.558** |

**Table 4: Post popularity prediction on item-specific setting.**

| | **Textual Features** | | **Visual Features** | | | **Multimodal** |
|---|---|---|---|---|---|---|
| | W2V | Sentiment | CNN-Pool5 | Concept | Sentiment | **Late Fusion** |
| PKNN | 0.195 | 0.101 | 0.211 | 0.132 | 0.178 | 0.226 |
| PSVR | 0.492 | 0.315 | 0.501 | 0.425 | 0.446 | 0.534 |
| PRSVM | 0.486 | 0.271 | 0.461 | 0.427 | 0.439 | 0.503 |
| PBG | 0.489 | 0.294 | 0.487 | 0.419 | 0.440 | 0.435 |
| PCFM(ours) | **0.511** | **0.319** | **0.545** | **0.459** | **0.491** | **0.576** |

optimal rank correlation. We have considered $k \in \{8, 16, 32, 64\}$ as the dimension of the factorization for learning the parameter vector $\Theta$. We found that on the validation set best results are obtained with $k = 64$. We consider 100 iterations as stopping criteria in our algorithm.

**Post popularity** As the measure of post popularity we use the number of likes it received as used in [12, 18]. We find that the number of likes follows a power law distribution, where the majority of posts receive little or no likes and the minority of them receive a high number of likes. To deal with the large variation in the number of likes, we apply the log function as used in [12] to make it resemble a Gaussian distribution of the like counts.

**Evaluation metric** After predicting the popularity of each post at test time we compute the Spearman's rank correlation between the prediction and ground truth which returns a value between [-1, 1]. A value close to 1 corresponds to perfect positive correlation.

## 4.3 Experiments

*Experiment 1:* ***Baseline creation*** In order to compare our multimodal context-aware recommender system with state-of-the-art methods in popularity prediction, we consider the following baselines:

- *Baseline 1*: **Popularity by KNN(PKNN)**. Inspired by [17] which showed that similar images share similar tags, we hypothesize in this baseline that similar posts tend to obtain similar popularity independent of the user which shared the post and the item the post is about. We estimate the popularity of a post in test time by averaging the popularity of the top k-nearest neighbors (k=10 in this paper) posts in the train set. We use euclidean distance as a metric for computing the similarity score.
- *Baseline 2*: **Popularity by SVR(PSVR)**[18]. This baseline considers the popularity prediction as a ranking problem and trains a support vector regressor over all posts of the train set without considering the interaction between users which shared posts and the items of the posts.
- *Baseline 3*: **Popularity by Rank-SVM(PRSVM)**. This baseline also considers the problem of popularity prediction as a ranking problem and uses a Rank-SVM [11] for ranking the posts at test time.
- *Baseline 4*: **Popularity by Bipartite Graph(PBG)** [8]. This baseline models the popularity prediction of posts using a bipartite graph which only considers the interaction of users with items without considering the interaction between the aggregation of all users and the aggregation of all items.
- *Baseline 5*: **Popularity by Factorization Machine (PFM)** [13]. As context-unaware baseline we use FM where only the user and item variables are used, without considering the context constraints inside the objective function, for generating the feature vectors. This is equivalent to matrix factorization with bias terms.

We evaluate all these baselines on two settings: *user-specific,* and *item-specific* using different visual and textual features and the fusion with a late fusion by average operator.

*Experiment 2:* ***Ours versus the baselines*** In this experiment we evaluate the effect of our proposal, Popularity prediction by single modal Context-aware Factorization Machines, *PCFM*, and Popularity prediction by Multimodal Context-aware Factorization Machines, *PMCFM.* In *PCFM* method, we only consider one modality, visual or textual, as contextual information of posts in formula (6). In fact we use three variables, *U, I,* with one of the variables *VC* or *TC* in section 3.3. We use all contextual information explained in section 4.2 in our *PCFM* individually and report the result. Moreover we report the result of fusing all modalities by average operator as a late fusion. In the *PMCFM* method, we select the best visual context, $VC_{best}$, and textual context, $TC_{best}$, based on the result of the *PCFM* method. Then we use all possibilities for variable $var = \{U, I, VC_{best}, TC_{best}\}$ for learning the parameter vector $\Theta$ in formula (6). At the end, we compare the result of our proposal with all baselines.

*Experiment 3:* ***Selection of posts from an offline collection of a specific user or a specific item*** In this experiment we evaluate

**Table 5: Comparison of our multimodal context-aware recommender against context-aware and context-unaware baselines for popularity prediction on both settings of our dataset. Our method outperforms the others for predicting the popularity of post in both settings.**

| | Dataset | |
|---|---|---|
| Method | user-specific | item-specific |
| PFM | 0.410 | 0.466 |
| PCFM (ours) | 0.558 | 0.576 |
| PMCFM (ours) | 0.592 | 0.610 |

the effect of our method versus all baselines for selecting those images from an off-line collection of images of a specific user and specific items which are expected to get a high number of likes. For this purpose we use the *user-specific*, and *item-specific* settings of our datset in test time and use only the visual data. We use the visual features per image in the test set and apply all the pre-trained models, baselines and our proposals, of experiment 1 and experiment 2 to compute a popularity score (Note since we are using only visual features we report the result of our *PCFM*). Then, we rank the images for a specific user from the *user-specific* setting based on their popularity score and select the top 10 images as most promising for sharing in social networks. We evaluate the selection, by defining the popularity ratio as average number of matches between images selected by methods and images from the ground truth ranking. The value of the popularity ratio shows the quality of the approach for selecting images to be shared in social networks. Proximity of the popularity ratio to 1 indicates a better image selection. We repeat this procedure on *item-specific* to compute the popularity score of selecting an image for a specific item.

# 5 RESULTS

## 5.1 Baseline creation

We report the popularity prediction performance of all baselines in both settings of our dataset, *user-specific* and *item-specific*, in Table 3 and Table 4 respectively. As we can see PKNN is the weakest baseline and its best correlation is only about 0.206 in *user-specific* and 0.226 in the *item-specific* setting as it ignores the abundant contextual data of popularity, just similarity in content. The result of PFM in Table 5 reaches 0.410 and 0.446 in the respective settings which is popularity modeling based on user-item matrix decomposition without using any contextual data. That suggests the effect of modelling popularity using a recommender. Besides, the popularity models of PSVR and PRSVM generally outperform PFM. It implies the effect of visual and textual contextual information on popularity prediction. The results in Table 3 and Table 4 depict the efficiency of PSVR in comparison with the other baselines where the results of rank correlation reach to 0.511 and 0.534 in *user-specific* and *item-specific* respectively.

## 5.2 Ours versus baselines

In Table 3 and Table 4, we give the prediction performance of our proposal, PCFM, and compare it with the baselines using different modalities. As we can see in all columns of both tables, our method achieves the best performance using different contextual features and multimodal late fusion. The results in both settings of our dataset show the relative improvements of our approach over the best baseline model, PSVR, is about 10% and 8% on *user-specific* and *item-specific* setting respectively. It shows the importance of considering user-item interaction with contextual features of posts in modeling popularity prediction. The results in Table 5 show the 36% and 29% relative improvement of comparing PCFM with PFM in *user-specific* and *item-specific* setting respectively. It suggests the user-item matrix decomposition with considering contextual data is more adequate to utilize in the popularity prediction model. The results in Table 3, and Table 4 also show that our PCFM method reaches the best results on both settings using W2V as a textual context and CNN-Pool5 as a visual context. We consider these two features, as the best features, for making our PMCFM model.

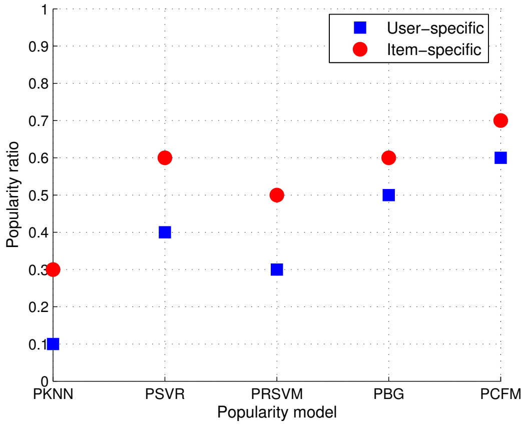

**Figure 3: The result of image selection based on all popularity models. Our method outperforms the others for selecting the popular images for specific-user and -item.**

We report the result of PMCFM, in Table 5. The result of rank correlation using PMCFM reach 0.592 and 0.610 in the *user-specific* and *item-specific* setting respectively. It shows 7% and 6% relative improvement in comparison with our PCFM on both settings. It emphasizes we should use both visual and textual context of posts.

The results in Table 3 , 4, and 5 show the performance of popularity prediction on *item-specific* is slightly better than *user-specific*. This indicates that the prediction on images belonging to different items with different contents is more difficult than the images which show contents belonging to the same items.

The results of experiment 2 on both settings of our dataset show that, in general, the post popularity prediction accuracy increases when considering the user-item interaction inside the model. Moreover, keeping the interaction of users with items and multimodal contextual information of posts achieves the best performance and give higher relative rank correlation than the other methods.

## 5.3 Image Selection for a specific-user or -item

We display the results of experiment 3, using fusion of all visual features for image selection, in Figure 3. The results demonstrate the effectiveness of our proposal, PCFM, against all baselines for selecting images, which have the potential of getting more likes, from a collection of images of a specific-user or -item. When we

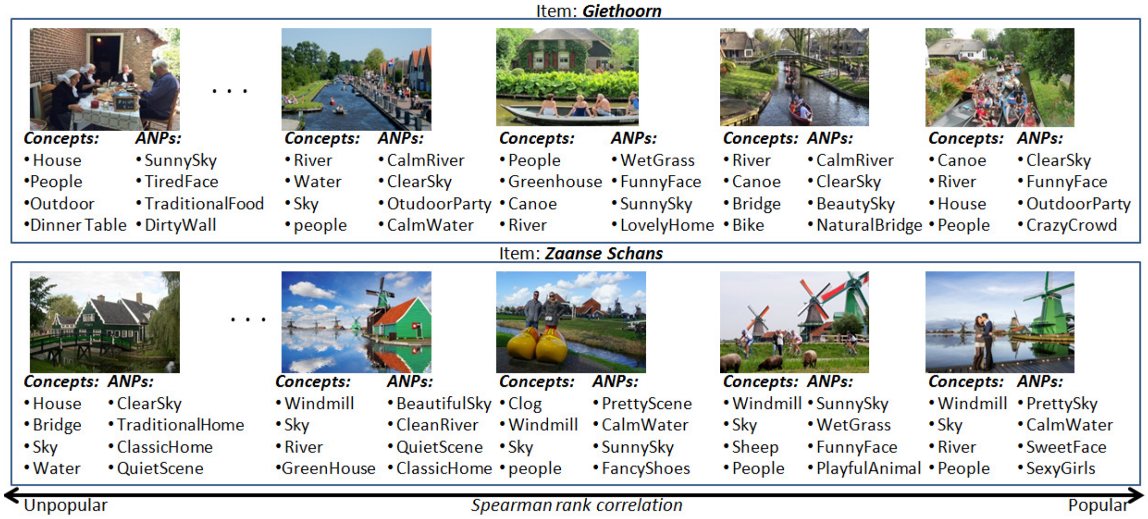
Item: Giethoorn
Concepts: • House • People • Outdoor • Dinner Table
ANPs: • SunnySky • TiredFace • TraditionalFood • DirtyWall
Concepts: • River • Water • Sky • people
ANPs: • CalmRiver • ClearSky • OtudoorParty • CalmWater
Concepts: • People • Greenhouse • Canoe • River
ANPs: • WetGrass • FunnyFace • SunnySky • LovelyHome
Concepts: • River • Canoe • Bridge • Bike
ANPs: • CalmRiver • ClearSky • BeautySky • NaturalBridge
Concepts: • Canoe • River • House • People
ANPs: • ClearSky • FunnyFace • OutdoorParty • CrazyCrowd
Item: Zaanse Schans
Concepts: • House • Bridge • Sky • Water
ANPs: • ClearSky • TraditionalHome • ClassicHome • QuietScene
Concepts: • Windmill • Sky • River • GreenHouse
ANPs: • BeautifulSky • CleanRiver • QuietScene • ClassicHome
Concepts: • Clog • Windmill • Sky • people
ANPs: • PrettyScene • CalmWater • SunnySky • FancyShoes
Concepts: • Windmill • Sky • Sheep • People
ANPs: • SunnySky • WetGrass • FunnyFace • PlayfulAnimal
Concepts: • Windmill • Sky • River • People
ANPs: • PrettySky • CalmWater • SweetFace • SexyGirls
Unpopular
Spearman rank correlation
Popular

**Figure 4: The result of selecting image for two items: *Giethoorn* and *Zaanse Schans* based on our proposal, from popular to unpopular. As we can see the top selected images, popular, show the important and relevant concepts related to the items.**

request to select 10 images we reach 0.6 and 0.7 accuracy in popularity ratio using *PCFM*, on user- and item-specific respectively. It means our proposal can select accurately 6 and 7 images out of 10 images in both settings. Whilst using *PKNN, PSVR, PRSVM* and *PBG* the results reach 0.1, 0.4, 0.3, and 0.5 in the user-specific setting and 0.3, 0.6, 0.5, and 0.6 in the item-specific setting.

We observe from this experiment that the presence of *Concepts* and *Visual Sentiments* for all specific-user and -item has a positive effect on the predicted popularity. We further investigate which particular *Concepts* and *ANPs* correlate most with the number of likes for a specific item. To evaluate the correlation of *Concepts* and *ANPs* with the popularity of items, we follow this procedure: i) For each item we select the 100 most popular images of the train set, based on the number of likes. ii) We select the top 1000 *Concepts* and *ANPs* on each selected image of an item based on their probability. iii) We sort *Concepts* and *ANPs* based on their frequency of occurrence in selected images and consider the top *Concepts* and *ANPs* as important features for the item. For example for item *Giethoorn*, a touristic place in the Netherlands, we find the existence of *Concepts* such as *Canoe, River, Bridge, Biking,* and *Traditional House* are most important. Also we find those *ANPs* related to scenes such as *sky, water,* and *face* of people are important.

We investigate the existence of the top *Concepts* and *ANPs* on selected image for two items, *Giethoorn* and *Zaanse Schans*. The results in Figure 4, highlight the important *Concepts* and *ANPs* in selected images for these items. As we can see those image selected with our proposal from test set as more popular for item *Giethoorn*, are those images which include concepts *Canoe, River, Bridge*. Less popular images are those images of *Giethoorn* which don’t show these important concepts. The results of experiment 3 show, in general, we have the ability of helping users, tourist organizations and brand companies in selecting the content they should share on their social media page.

## 6 CONCLUSION

In this paper we propose to consider the interaction between users and items for predicting the popularity of posts related to a specific-user and a specific-item in social media. Different from existing works which aggregate all user posts related to different items and ignore the preferences of individual users to the items, we present an approach which considers the preferences of individual users to the items for predicting the popularity of posts related to a specific-user and -item. We factorize the popularity of posts to the user-item-context, make a popularity tensor and use a multimodal context-aware recommender for predicting post popularity. Since we use a recommender for modeling the popularity of a post, we have the ability of simultaneously predicting the popularity of posts related to different items which are shared by a specific user and a post shared with different users for a specific item.

We study the behavior of our proposal by performing three experiments on a collection of user posts related to touristic places in the Netherlands crawled from Instagram. The results of experiment 1 and 2 demonstrate the effectiveness and power of our proposal versus the state-of-the-art methods in post popularity prediction. It shows a relative improvement in user-specific and item-specific post popularity prediction of 16% and 14% compared with the best baseline. Moreover, experiment 3 reveals that our proposal, where all visual features are fused, outperforms the other methods for selecting a set of off-line images of a specific-user and -item likely to become most popular.

We conclude that for predicting the popularity of a post for a specific-user and -item it is beneficial to consider the user-item-context interaction and construct a multimodal context-aware recommender for modeling the popularity of a post.

## ACKNOWLEDGMENTS

This research is supported by the Amsterdam Academic Alliance Data Science (AAA-DS) Program Award of the UvA and VU Universities.

## REFERENCES

[1] Younggue Bae and Hongchul Lee. 2012. Sentiment Analysis of Twitter Audiences: Measuring the Positive or Negative Influence of Popular Twitterers. *J. Am. Soc. Inf. Sci. Technol.* 63, 12 (2012), 2521–2535.

[2] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *MM*.

[3] Spencer Cappallo, Thomas Mensink, and Cees GM Snoek. 2015. Latent factors of visual popularity prediction. In *ICMR*.

[4] Heng-Yu Chi, Chun-Chieh Chen, Wen-Huang Cheng, and Ming-Syan Chen. 2016. UbiShop: Commercial Item Recommendation Using Visual Part-based Object Representation. *Multimedia Tools Appl.* 75, 23 (2016), 16093–16115.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

[6] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. 2015. Image popularity prediction in social media using sentiment and context features. In *MM*.

[7] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. 2015. Image Popularity Prediction in Social Media Using Sentiment and Context Features. In *MM*.

[8] Xiangnan He, Ming Gao, Min-Yen Kan, Yiqun Liu, and Kazunari Sugiyama. 2014. Predicting the Popularity of Web 2.0 Items Based on User Comments. In *SIGIR*.

[9] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. 2011. Predicting Popular Messages in Twitter. In *WWW*.

[10] Instagram Inc. 2016. Instagram Developer API. (2016). https://www.instagram.com/developer/ Online; accessed: 15-July-2016.

[11] Thorsten Joachims. 2002. Optimizing Search Engines Using Clickthrough Data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[12] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. 2014. What Makes an Image Popular?. In *WWW*.

[13] Deguang Kong, Chris Ding, and Heng Huang. 2011. Robust Nonnegative Matrix Factorization Using L21-norm. In *CIKM*.

[14] Cheng Li, Yue Lu, Qiaozhu Mei, Dong Wang, and Sandeep Pandey. 2015. Click-through Prediction for Advertising in Twitter Timeline. In *KDD*.

[15] Amsterdam Marketing. 2016. Amsterdam Open Data - Attractions. (2016). https://data.amsterdam.nl/dataset/attracties Online; accessed: 15-July-2016.

[16] Masoud Mazloom, Amirhossein Habibian, Dong Liu, Cees G. M. Snoek, and Shih-Fu Chang. 2015. Encoding Concept Prototypes for Video Event Detection and Summarization. In *ICMR*.

[17] Masoud Mazloom, Xirong Li, and Cees G. M. Snoek. 2016. TagBook: A Semantic Video Representation without Supervision for Event Detection. *TMM* 18, 7 (2016), 1378–1388.

[18] Masoud Mazloom, Robert Rietveld, Stevan Rudinac, Marcel Worring, and Willemijn Van Dolen. 2016. Multimodal Popularity Prediction of Brand-related Social Media Posts. In *MM*.

[19] Philip J McParlane, Yashar Moshfeghi, and Joemon M Jose. 2014. Nobody comes here anymore, it's too crowded; Predicting Image Popularity on Flickr. In *ICMR*.

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

[21] Xiang Niu, Lusong Li, Tao Mei, Jialie Shen, and Ke Xu. 2012. Predicting Image Popularity in an Incomplete Social Media Community by a Weighted Bi-partite Graph. In *ICME*.

[22] Steffen Rendle. 2010. Factorization Machines. In *ICDM*.

[23] Steffen Rendle, Zeno Gantner, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2011. Fast Context-aware Recommendations with Factorization Machines. In *SIGIR*.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. In *CVPR*.

[25] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.* 61, 12 (2010), 2544–2558.

[26] Zhi Wang, Wenwu Zhu, Xiangwen Chen, Lifeng Sun, Jiangchuan Liu, Minghua Chen, Peng Cui, and Shiqiang Yang. 2013. Propagation-based Social-aware Multimedia Content Distribution. *ACM Trans. Multimedia Comput. Commun. Appl.* 9, 1s (2013).

[27] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, and Tao Mei. 2016. Time Matters: Multi-scale Temporalization of Social Media Popularity. In *MM*.

[28] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. 2016. Unfolding Temporal Dynamics: Predicting Social Media Popularity Using Multi-scale Temporal Decomposition. In *AAAI*.

[29] Chun-Che Wu, Tao Mei, Winston H. Hsu, and Yong Rui. 2014. Learning to Personalize Trending Image Search Suggestion. In *SIGIR*.