Prediction of image popularity over time on social media networks

Conference Paper · October 2016

DOI: 10.1109/CT-IETA.2016.7868253

CITATIONS
10

READS
601

3 authors, including:

Khaled Almgren
University of Bridgeport
8 PUBLICATIONS **90** CITATIONS
SEE PROFILE

Minkyu Kim
13 PUBLICATIONS **109** CITATIONS
SEE PROFILE

Prediction of Image Popularity Over Time on Social Media Networks

Khaled Almgren, Jeongkyu Lee
Dep of Computer Science and Engineering
University of Bridgeport
Bridgeport, CT, 06614
kalmgren@my.bridgeport.edu, jelee@bridgeport.edu

Minkyu Kim
ASML
77 Danbury Rd, Wilton, CT, 06877
minkyu.kim@asml.com

Abstract— **This paper introduces a new problem that investigates image popularity on social networks. It is referred to as the stability of image popularity. It focuses on whether the popularity of a certain image on social networks will change over time or not. In order to investigate what drives the changes of popularity, we employ three features of social media data, i.e., social context, image semantics, and an image's early popularity. To predict the changes of images popularity, we collected 69,000 images from Instagram and kept track of their popularity for one month. For the popularity measure, we use Pareto principle. Using a Gaussian Naive Bayes classifier, we predict the changes of image popularity, and the results show that the semantic feature is the only one that predicts the changes in the popularity of an image over time.**

Keywords— popularity prediction; social media images; captions; time-sensetive popularity;

I. INTRODUCTION

Social networks have changed the way how information spread. Basically, users in social media networks propagate information in the form of posts using interactions. For example, users can share a piece of information on Twitter in the form of tweets, and tweets can be transferred using retweets. On Instagram, information can be shared in the form of images and propagated based on likes. This phenomenon is referred as information diffusion.

The power of social networks have attracted researchers from academia and business to conduct a vast amount of research to understand the phenomena of information diffusion, which introduces several research topics, including image popularity prediction.

Previous studies on this topic predicted the overall image popularity without considering the changes of popularity over time [3–6, 8, 10, 11, 14, 20–22]. However, image popularity is not necessarily constant. Instead, it can be increased or decreased along the time.

In order to address this, we introduce a new problem that focuses on predicting the changes in an image popularity over time. This problems is referred as the stability of an image popularity. It is defined as follows: At time t_i an image i popularity is defined as one of two states, i.e., 0 or 1, where 1 indicates popular and 0 indicates unpopular in terms of the number of likes that image i receives. At t_{i+1}, image i popularity state can change or stays the same based on the number of likes that image i receives at t_{i+1} and so on. If the popularity state of image i stays unchanged at all times, then image i popularity is stable. Otherwise it is changing.

In order to predict the popularity stability, social context, image semantics, and early popularity are investigated. For social context, several researchers observed a correlation between users' popularity and their posts popularity [4, 6, 10]. Therefore, we adopt the number of followers of the user who uploads an image because it represents the popularity of the user. However, not all images from the same user become popular. Therefore, we examine other features that can drive image popularity such as image content. For the image content, we employ our previous approach [1]. The approach is based on natural language processing and clustering techniques that are used to extract the semantics from images using their captions. A strong correlation between early and future popularity is reported in [18]. This observation is adapted us to investigate the effect of early popularity of an image on the stability of its popularity.

Predicting an image's popularity can be very useful for marketing; if we can predict that an image will keep or lose its popularity over time, we can help companies target their market effectively. The proposed work is evaluated by using Instagram (a multimedia social network with 400 million users) [9].

The contributions of this paper are summarized below:

1. A new problem, i.e., popularity stability, is defined, which focuses on the changes of image popularity over time.
2. A new study is initiated, which investigates what may drive an image's popularity to change over time.
3. To the best of our knowledge, this is the first study that report that image content is the only feature that can predict the changes of the image's popularity over time.

The rest of this paper is organized as follows: the related work is briefly discussed in Section II. In Section III, the popularity measurement is explained. The features are discussed in Section IV. Our experiments and results are

978-1-5090-0799-8/16/$31.00 ©2016 IEEE

shown in Sections V and VI respectively. We conclude in Section VII.

II. Related work

We review the recent studies on the topic of predicting the popularity of posts, and categorize them in terms of the data type, approach and problem type.

A. Social media data type

B.

In this category, we classify the research papers based on the social media data type that they are focused on, which are text, and multimedia with a focus on images.

1) Text-based social media data: Yu et al. tried to predict how many times a tweet get retweeted using user, text, and temporal features [21]. Hong et al. predicted the popularity of tweets using tweet content, topical information of tweets, users, and temporal features [8]. Zaman et al. focused on predicting wither a tweet will be retweeted or not using interaction patterns between users, users information, and tweet content [22]. They all measured popularity using the number of retweets [8,21-22].

2) McParlane et al. predicted the popularity of images on Flickr using image content, image content, and user information [11]. For the image content, they classified the images according to a set number of scenes. They measured popularity using the number of comments and views. Khosla et al. predicted how many times an image get viewed on Flickr using image content, and social context [10]. Cappallo et al. predicted the popularity of images on Flickr and Twitter using images content [4]. They considered content from both popular and unpopular images. They measured popularity using the normalized number of views. Can et al. predicted the popularity of images posted on Twitter and Flickr using hash tags, user information and image low level and high level features, such as color [3]. They measured the popularity of images on twitter based on the number of favorites and retweet, and number of views and comments on Flickr. Yamaguchi et al. employed social, content, and text features to predict the popularity of images on Chictopia [20]. They measured popularity based on the number of votes. Totti et al. classified the popularity of images using aesthetical, semantic, and social features on Pinterest [19]. They measured popularity using the number of repins. Fiolet classified the popularity of images on Instagram by ranking the popularity of images using user and image information [5]. They measured popularity based on the number of likes. Niu et al. ranked the popularity of images using network based featured on Flickr [14]. They used the number of views as a popularity measurement. Gelli used visual sentiments, image content, and context features to predict the normalized number of views of images on Flickr [6].

C. Approach

The approaches represent the algorithms or models that researchers use to perform the prediction. The related work either used a learning model, such as support vector machine, or a none-learning model, such as network measures.

1) Learning Models: Yu et al. employed a logistic regression model [21]. Hong et al. use a logistic regression classifier [8]. Zaman et al. employed a collaborative filtering model [22]. Several papers used support vector machine for image popularity prediction [11]. Khosla et al. employed a regression model based on support vector machine [4,6,10]. Can et al. used regression model based on linear regression, support vector machine, and random forest [3]. Yamaguchi used regression analysis [20]. Totti et al. used random forest [19].

2) None-Learning Models: Fiolet simply ranked the images based on different features without using any prediction model [5]. Niu et al. employed a weighted bipartite graph model [14].

D. Problem Type

Recent works have formalized the problem of popularity prediction in three ways, i.e., classification, retrieval, and regression. In regression, researchers tried to quantify the popularity of posts, where in classification, they classified the popularity to a set of classes, such as popular or not popular. In retrieval, researchers ranked the images from most to least popular.. Yu et al. formalized the problem as regression one [3,21]. Several papers formalize the problem as a classification problem [8,11,19,22]. Several papers formalized the problem as a retrieval one [4-6,10,14,20].

III. Popularity measurement

Webster's dictionary defines popularity as ``*the state of being liked, enjoyed, accepted, or done by a large number of people* [10]. " This can be reflected on social networks through users' actions. Users on Instagram can like an image by pressing on the *heart button* as shown in Figure 2. Intuitively, the higher number of likes that an image receives, the more popular it becomes. Therefore, the number of likes is adopted as the popularity measurement as used in [1]. Moreover, the number of likes is classified to either low or high, where low is unpopular and high is popular. We adopt the Pareto principle (80%-20%) to compute the popularity threshold using the number of likes as used in [1,11]. An image that receives a number of likes that is greater or equal to this threshold is considered popular. The thresholds of popularity are: likes greater or equal than 49 for the first hour, 69 for the next day, 75 for the next week, and 76 for the next month.

For example, let's assume that image *i* gets 49 likes during the first hour of sharing the image, thus, it is considered as popular. After the first day, the number of likes on image *i*

increases to 70, so it will stay popular for the first day. Then, after the first week, if the number of likes is less than 75, image image *i* will be unpopular. After one month, if the number of likes stay less than the threshold, image *i* will stay unpopular. Overall, image *i* started out popular, then it lost its popularity after 1 week. This is one scenario that represent a changing popularity.

Figure 1 shows the distribution of the number of images with respect to the number of likes in the first hour, next-day, first week, and first month. Both axis are log scaled. The X axis represents the normalized number of likes by the popularity threshold for each time period, therefore, the overlapped line represents the normalized popularity thresholds. The Y axis represents the normalized number of images by the maximum number of images.

IV. FEATURES

In this section, we discuss the three features, i.e., image semantics, an image's early popularity, and the user's information.

A. Social Context

Previous works showed that users popularity is correlated with their image popularity [3, 5, 8, 9, 17]. Therefore, in this work, we use users information to predict the changes in the popularity of their uploaded images. To normalize the number of followers, we take the 10th log of the number of followers, and then normalize it by the maximum number of followers in our dataset as adopted in [1]. This is computed using the following equation:

$$Sf_i = \log_{10}(\#\,fol + 1) \div Max(\#\,fol), \quad (1)$$

where fol is the number of followers.

B. Image Semantics

Several studies on this topic showed that image content has an impact on it popularity [4,5,8]. However, in their analysis, they only consider the visual proprieties of an image, such as color, sentiment, textures, and visual scenes. Even so, these features reflect the content, but does it really reflect the meaning of the content. In this paper, we argue that it does because of the subjectivity of semantics. Oglesbee states that ``*Looking at a picture without a caption is like watching television with the sound turned off*'' [12].

Understanding the meaning of images can be challenging. Therefore, photographers use captions to express the meaning of images. Captions are a small description of images that are usually placed under the images (See Figure 2). Therefore, multimedia social networks, such as Instagram, support captions.

In this paper, we analyze the effect of an image's semantics on its popularity. For that, we use captions to extract the semantic of an image using natural language processing and clustering techniques. The approach is explained in [1]. To briefly describe it, the first step is to pre-process the captions to generate keywords, then use Word2vec to convert to the keywords to numerical forms that can be used to detect similarity between keywords. Finally, by employing k-means, we generate keyword vector that can represent the multiple semantics of an image (Figure 3 illustrates the approach).

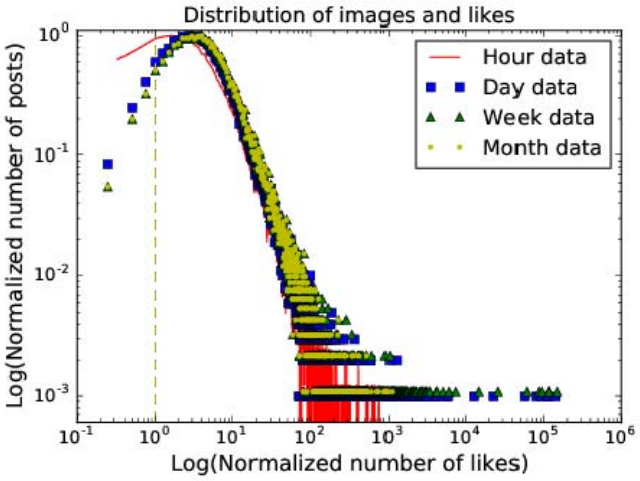

Figure 1: The graph represents the distribution of images and likes for the first hour, next-day, first week, and first month. The line represent the overlapped normalized popularity thresholds. The x-axis is the normalized number of likes while the y-axis is the normalized number of images.

C. Early Popularity

Szabo et al. reported a strong linear correlation between early and future popularity on Youtube and Digg [15]. We observed similar trends from the distribution of images and likes over different time frames as shown in Figure 1. This means that there is a possibility that early information about the popularity of an image is correlated with its future popularity. Based on these observations, we adopt the early popularity to predict the popularity stability. On Instagram, we can collect data with respect to an image's popularity, i.e., number of likes during the first hour when the image is uploaded. We then employ the popularity threshold to classify the early popularity. This feature is a binary variables that represents the popular images as 1 and 0 for unpopular as adopted in [1]. The popularity variable, i.e. *EP*, is computed based on the Pareto principle threshold as follows:

$$EP_i = \begin{cases} 1 : if\ (likes \geq threshold \\ else : 0 \end{cases} \quad (2)$$

V. EXPERIMENTAL SETUP

A. Datasets

The dataset is crawled using the Instagram API[1]. Two methods on Instagram are available to crawl images. The first one retrieves recent images using given users IDs, while the second method retrieves images based on a given geographical location. The first method was used to make sure the data is completely random. This approach requires users IDs. We randomly selected more than 1,000,000 IDs. Using these IDs, we triggered the Instagram API to check whether these users

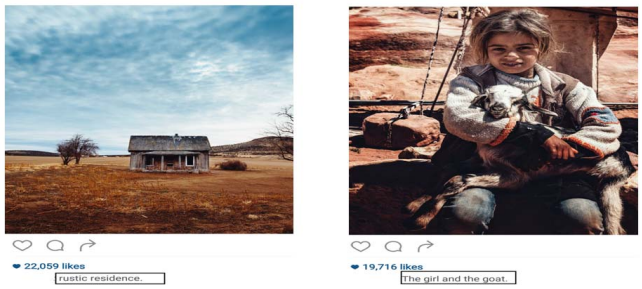

Figure 2: Two images retrieved from Instagram with their captions: the image on the left is described as "reustic house", while the image on the right is described as "the girl and the goat"

are private or public since users have the choice to make their profiles publicly available. We found 149,520 users who are public. However, among these users, there were 89,093 users who shared at least ten images. We used these users because they are active to retrieve two random datasets containing images that were uploaded during the first hour when we were collecting the data.

For our experiments. we retrieved 69,000 images. However, after preprocessing, we have 51,647 images. Set 1 contains 39,302 images, while Set 2 has 12,345 images. The first set was retrieved between January 2015 and February 2016, while set 2 was retrieved between March 2016 and April 2016. Set 1 is used for training and testing is performed on Set 2. These images have received 16,331,397 likes. Instagram does not provide the likes timestamps, therefore, we tracked the number of likes for each image using its ID for three periods: the next-day, after the first week, and after the first month. The dataset is available at [2].

B. Evaluation

For evaluating the accuracy of our model, we adopted sensitivity (Sen), i.e., true positive rate. Sensitive is widely used in the case of imbalanced data [7]. We were interested in predicting the changes of popularity, which is a phenomenon that happens rarely, thus, we adopt sensitivity. The sensitivity is computed as follow:

$$Sen = \frac{Number\ \_of\ \_true\ \_positives}{Number\ \_of\ \_images\ \_with\ \_changing\ \_popularity} \quad (3),$$

where true positives are images that their popularity had changed over time, and are correctly identified as changing popularity. A Gaussian Naive Bayes model is trained to predict the popularity stability of images using Set 1 and testing is performed on Set 2 to evaluate our model.

C. System Implementation

Gaussian Naive Bayes are implemented using Sklearn [16]. We chose Gaussian Naive Bayes because it computes the probability instead of distance. It also outperformed other classifiers, including SVM, and Random Forest.

VI. RESULTS

In this section, we discuss the impact of the three features on the stability of the popularity by conducting several experiments.

A. What makes image popularity change over time

Based on the analysis of our dataset, we see that images who usually start popular stay popular, and images that start unpopular stay unpopular. However, this is not the case all the times. We observed that the popularity of many images had changed over time. In this experiment, we investigate what may drive such a behavior.

We employ the three features discussed earlier to predict the stability of popularity. As shown in Table 1, we see that early popularity and user information cannot predict the changes of popularity, even so, they are observed to be important in predicting image popularity. On the other hand, image semantics can predict the popularity changes with a sensitivity rate of 0.34, which shows that the semantics of an image has an effect on it popularity as time passes.

B. What makes image popularity stable over time

In this experiment, we investigated what makes the popularity of an image stay stable over time (Note that the sensitivity matrix is updated to images with stable popularity to reflect the accuracy of this experiment).

We found that social context and early popularity are perfectly linked with stable popularity with a perfect sensitivity rate of 1.0. For the early popularity result, we hypothesis that this can be because popularity may be saturated after the first hour. Moreover, the social context result indicates that a user's popularity can make the popularity of his/her images stable over time. An image's semantics is also highly correlated with stable popularity with a sensitivity rate of 0.76, which also shows that the content of an image can make the image keep its popularity for a long time.

VII. CONCLUSION

In this paper, a new research problem was defined, i.e., popularity stability. We employ an image's early information to predict the stability of its popularity. The early information contains three features, i.e., social context, image semantics, and images' early popularity. Our results show that an image's semantics perform better in predicting the images

[1] https://www.instagram.com/developer
[2] http://www1bpt.bridgeport.edu/~jelee/sna/pred.html

with changing popularity, while the user's information and images early popularity perform better in predicting images with stable popularity.

For our future work, we will focus more on optimizing our semantics feature results since it is the only feature that shows potential for predicting the changes in the popularity of an image. We will further investigate other features that can cause such a phenomena. By making the Instagram dataset available, we hope that other researchers engage in solving these challenging problems.

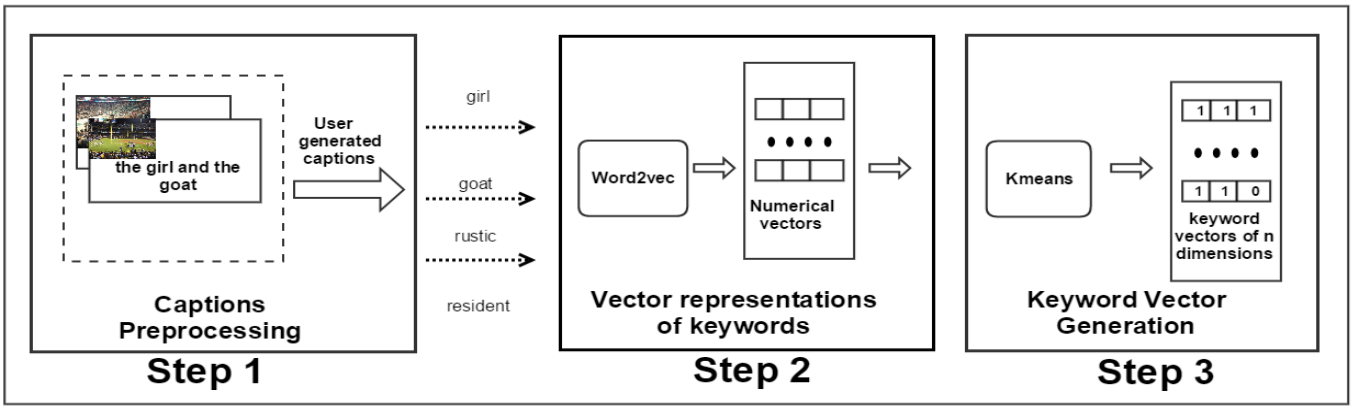

Figure 3: The three-step approach to extract and represent images semantics

Popularity Status\ Features	User	Semantic	Early Popularity
Changing	0.0	0.34	0.0
Stable	1.0	0.76	1.0

Table 1: Gaussian Naive Bayes Sensitivity rates for stable and changing popularity for each feature

REFERENCES

[1] Almgren, K., Lee, J., & Kim, Minkyu. (2016, August). Predicting the Future Popularity of Images on Social Networks. In The 3rd Multidisciplinary International Social Networks Conference (MISNC). ACM. (Accepted)

[2] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. " O'Reilly Media, Inc.".

[3] Can, E. F., Oktay, H., & Manmatha, R. (2013, October). Predicting retweet count using visual cues. In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (pp. 1481-1484). ACM.

[4] Cappallo, S., Mensink, T., & Snoek, C. G. (2015, June). Latent Factors of Visual Popularity Prediction. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (pp. 195-202). ACM.

[5] Fiolet, E. (2014, July). Analyzing Image Popularity on a Social Media Platform (Master Thesis). Retrieved from http://dare.uva.nl/cgi/arno/show.cgi?fid=544699

[6] Gelli, F., Uricchio, T., Bertini, M., Del Bimbo, A., & Chang, S. F. (2015, October). Image Popularity Prediction in Social Media Using Sentiment and Context Features. In Proceedings of the 23rd Annual ACM Conference on Multimedia Conference (pp. 907-910). ACM.

[7] Honest, H., & Khan, K. S. (2002). Reporting of measures of accuracy in systematic reviews of diagnostic literature. BMC health services research, 2(1), 1.

[8] Hong, L., Dan, O., & Davison, B. D. (2011, March). Predicting popular messages in twitter. In Proceedings of the 20th international conference companion on World wide web (pp. 57-58). ACM.

[9] Instagram. (2016) Stats. https://www.instagram.com/press/?hl=en

[10] Khosla, A., Das Sarma, A., & Hamid, R. (2014, April). What makes an image popular?. In Proceedings of the 23rd international conference on World wide web (pp. 867-876). ACM.

[11] McParlane, P. J., Moshfeghi, Y., & Jose, J. M. (2014, April). Nobody comes here anymore, it's too crowded; predicting image popularity on flickr. In Proceedings of International Conference on Multimedia Retrieval (p. 385). ACM.

[12] Merriam-Webster (2011) Popularity. http://www.merriam-webster.com/dictionary/popularity

[13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

[14] Niu, X., Li, L., Mei, T., Shen, J., & Xu, K. (2012, July). Predicting image popularity in an incomplete social media community by a weighted bi-partite graph. In Multimedia and Expo (ICME), 2012 IEEE International Conference on (pp. 735-740). IEEE.

[15] Oglesbee, L. 1998.' Captions. Looking at a picture without a caption is like watching television with the sound turned off', Communication: Journalism Education Today 32,2: 2-6.

[16] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825-2830.

[17] Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora.

[18] Szabo, G., & Huberman, B. A. (2010). Predicting the popularity of online content. Communications of the ACM, 53(8), 80-88.

[19] Totti, L. C., Costa, F. A., Avila, S., Valle, E., Meira Jr, W., & Almeida, V. (2014, June). The impact of visual attributes on online image diffusion. In Proceedings of the 2014 ACM conference on Web science (pp. 42-51). ACM.

[20] Yamaguchi, K., Berg, T. L., & Ortiz, L. E. (2014, November). Chic or social: Visual popularity analysis in online fashion networks. In Proceedings of the ACM International Conference on Multimedia (pp. 773-776). ACM.

[21] Yu, H., Bai, X. F., Huang, C., & Qi, H. (2015). Prediction algorithm for users Retweet Times.

[22] Zaman, T. R., Herbrich, R., Van Gael, J., & Stern, D. (2010, December). Predicting information spreading in twitter. In Workshop on computational social science and the wisdom of crowds, nips (Vol. 104, No. 45, pp. 17599-601). Citeseer.