Predicting the Popularity of Online Content with Group-specific Models

Qi Cao, Huawei Shen, Hao Gao, Jinhua Gao, Xueqi Cheng
{caoqi, gaohao, gaojinhua}@software.ict.ac.cn, {shenhuawei, cxq}@ict.ac.cn
CAS Key Laboratory of Network Data Science and Technology, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, China

ABSTRACT

Predicting the popularity of online content is highly valuable in many applications and has been studied for several years. However, existing models either work in population level—all messages are assumed to follow similar popularity dynamics, lacking flexibility to capture the intrinsic complexity of popularity dynamics, or work in individual level—the popularity dynamics of messages are independent of each other, failing to leverage other messages to improve prediction accuracy. In this paper, we propose a *divide and conquer* framework for popularity prediction. We first divide messages into groups, anticipating each group of messages follow similar popularity dynamics, and then, we train a group-specific model for the messages of each group. Experiments demonstrate that group-specific models improve the population-level models by about 30% and outperform state-of-the-art individual-level model.

Keywords

popularity prediction; divide and conquer; group-specific models

1. INTRODUCTION

The emergence of online social platforms, e.g., Twitter, Facebook, and Sina Weibo, allows everyone to become a producer of online content. Everyday, tens of millions of messages are generated on these platforms. Among them, a small amount of messages receive the most of attention, while a majority of messages get few attention. Being able to predict the popularity of online content is useful for both users and the owners of these platforms.

Existing methods for popularity prediction fall into two main categories: feature-based approaches and models based on generative process. The former typically works in *population* level, extracting various types of features (e.g., content, user, structural, and temporal features) and then training a single parametric model for all messages [1, 2]. These approaches actually assume that all messages follow an identical, at least similar, popularity dynamics with respect to observed features. Consequently, these approaches lack flexibility to capture the intrinsic complexity of popularity dynamics of messages. On the contrary, the latter attempts to model *individual* popularity dynamics of each message [3, 4]. These models generally assume that the popularity dynamics of messages are independent of each other and learn the message-specific parameters without leveraging the popularity dynamics of other messages.

In this paper, to circumvent the problem suffered by existing methods, we propose a *divide and conquer* framework for popularity prediction. First, we divide messages into groups, anticipating each group of messages follow similar popularity dynamics. Next, we train a parametric model for the messages of each group. Finally, the popularity of each message is predicted with group-specific models. We validate the effectiveness of the proposed framework by predicting the popularity of tweets.

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054242

2. MODELS

Suppose we have N messages, denoted by $\mathcal{M} = \{m^i\}(1 \leq i \leq N)$. For each message m^i, we use a cascade $\mathcal{C}^i = \{t^i_j\}$ to record the time elapsed between the original post and each retweet, where t^i_j corresponds to the jth retweet. The popularity R^i_t of message m^i up to time t is defined as the number of its retweets, i.e., $|\{t^i_j | t^i_j \leq t\}|$. The problem of popularity prediction is formulated as:

Popularity Prediction: Given the cascades in the observation time window $[0, T]$, it predicts the final popularity R^i_∞ of each message m^i.

2.1 Basic Models

We choose two representative population-level models, SH model and Pinto model, as our basic models.

• SH model [1]. This model is based on the observation that the future popularity of online content is linearly correlated with its early popularity. The final popularity is predicted by

$$\hat{R}^i_\infty = \alpha R^i_T.$$

• Pinto model [2]. It is an extension of SH model, replacing the single predictor, i.e., the cumulative popularity R^i_T, with multiple incremental popularity in equal-size time interval during the observation time window. The final popularity is predicted by

$$\hat{R}^i_\infty = \Theta \cdot \Delta R^i_T,$$

where $\Delta R^i_T = (R^i_{\Delta t}, R^i_{2\Delta t} - R^i_{\Delta t}, \ldots, R^i_T - R^i_{T-\Delta t})^T$, Δt is the length of time interval which can divide T exactly.

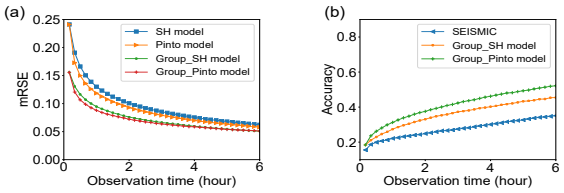
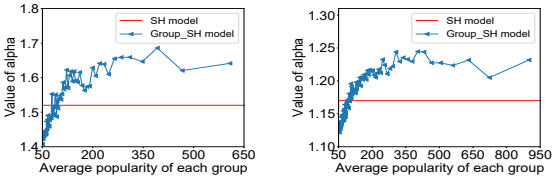


Figure 1: Popularity prediction performance.



(a) $T = 30$ minutes (b) $T = 6$ hours

Figure 2: Variation of model parameter for different groups.

2.2 The Divide and Conquer Framework

The proposed framework includes two stages. First, we divide messages into groups according to certain grouping strategy. For simplicity, we adopt the cumulative popularity in observation time window $[0, T]$ as grouping criterion. In other words, messages with the same popularity are divided into the same group. Now we have

$$\mathcal{G} = \{g^i\}(1 \leq i \leq N),$$

where $g^i \in \{1, 2, \ldots, G\}$ denotes the group ID of message m^i, and G is the number of groups.

Next, we train a parametric model for each group based on the above two basic models, obtaining two group-specific models:

- Group_SH model: $\hat{R}^i_\infty = \alpha_{g^i} R^i_T$;
- Group_Pinto model: $\hat{R}^i_\infty = \Theta_{g^i} \cdot \Delta R^i_T$.

Parameters of these two models are learned by

$$\min_{\alpha_g/\Theta_g} \sum_{i:g^i=g} f_{loss}(\hat{R}^i_\infty, R^i_\infty)(1 \leq g \leq G),$$

where f_{loss} denotes the loss function.

3. EXPERIMENTS

We now validate the proposed method using a Twitter dataset [4]. It contains 166,076 tweets, spanning from October 7 to October 21, 2011. Each tweet has at least 50 retweets within 7 days. We adopt the same way used in [4] to split the dataset, i.e., using the tweets from the first 7 days as training set and the remaining tweets as test set. The final popularity R_∞ of a tweet is approximated by R_{7days}. To prevent overfitting, we set the number of messages in each group larger than 500 (if not, merge the groups with similar popularity). The length T of observation time window ranges from 10 minutes to 6 hours with the incremental step being 10 minutes. The length of time interval Δt is set to be 10 minutes. Finally, we choose relative square error (RSE) as our loss function, *i.e.*, $f_{loss} = (\hat{R}_\infty/R_\infty - 1)^2$, and mean RSE (mRSE) over all messages as the evaluation metric.

3.1 Prediction Performance

Figure 1(a) compares our models with two population-level models. We can see that Group_SH and Group_Pinto model consistently reduce the prediction error of the corresponding basic models, reducing 30% and 29% respectively when only peeking 30 minutes into the future. The magnitude of error reduction decreases with the length of observation time window. This is attributed to the decrease of complexity of future popularity dynamics as we observe for a longer time.

We further compare our models with state-of-the-art individual level model, i.e., SEISMIC [4]. Since SEISMIC predicts an infinite popularity for some messages, mRSE is inappropriate for a fair comparison. Alternatively, we use accuracy for measurement, i.e., the fraction of messages that RSE is less than 0.01. As shown in Figure 1(b), Group_SH model and Group_Pinto model consistently outperform SEISMIC.

3.2 Analysis of Group-specific Models

To clarify why the divide and conquer framework works, we take Group_SH model as an example to show the variation of model parameters for different groups. Figure 2 shows the value of α with respect to the average cumulative popularity of tweets in each group. To avoid the dataset limitation that each tweet has at least 50 retweets, we only show the results on groups with the average popularity no less than 50. The high variance of α confirms that we do need group-specific models for popularity prediction.

4. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose a *divide and conquer* framework to predict the popularity of online content. Experiments conducted on a Twitter dataset demonstrate the effectiveness of the proposed framework. We also offer a comprehensive analysis to clarify why the proposed framework works. In the future, we will devote to finding better grouping strategy, using generative process to capture the popularity dynamics or using loss function as a guide.

5. ACKNOWLEDGMENTS

This work was funded by the National Basic Research Program of China (No. 2013CB329602) and the National Natural Science Foundation of China (Nos. 61472400, 61425016, 61433014). H. W. Shen is also funded by Youth Innovation Promotion Association CAS.

6. REFERENCES

[1] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[2] H. Pinto, J. M. Almeida, and M. A. Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *WSDM*, pages 365–374, 2013.

[3] H. W. Shen, D. Wang, C. Song, and A.-L. Barabási. Modeling and predicting popularity dynamics via reinforced poisson processes. In *AAAI*, pages 291–297, 2014.

[4] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *SIGKDD*, pages 1513–1522, 2015.