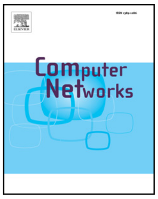
Computer Networks 140 (2018) 112–125

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet



Predicting the implicit and the explicit video popularity in a User Generated Content site with enhanced social features

Adele Lu Jia[a], Siqi Shen[b,c,*], Dongsheng Li[b,c], Shengling Chen[b,c,d]

[a] College of Information and Electrical Engineering, China Agricultural University, China
[b] School of Computer, National University of Defense Technology (NUDT), China
[c] Parallel and Distributed Processing Laboratory, NUDT, China
[d] Baidu Inc., China

ARTICLE INFO

Article history:
Received 31 January 2018
Revised 13 April 2018
Accepted 7 May 2018

Keywords:
User Generated Content (UGC) sites
Social features
Popularity prediction

ABSTRACT

User Generated Content (UGC) sites like YouTube are nowadays entertaining over a billion people. Identifying popular contents is essential for these giant UGC sites as they allow users to request contents from a potentially unlimited selection in an asynchronous fashion. In this work, we conduct an analysis on the popularity prediction problem in UGC sites and complement previous work with two new aspects, namely differentiating contents that attract a lot of attention and that users really appreciate, and leveraging built-in social features to predict the content popularity immediately upon publication.

To this end, we conduct an extensive measurement and analysis of BiliBili, a YouTube-like UGC site with enhanced social features including user following, chat replay, and virtual money donation. Based on datasets that contain over 2 million videos and over 28 million users, we characterize the video repository and the user activities, we analyze the video popularities, we propose graph models that reveal user relationships and high-level social structures, and we successfully apply our findings to build machine-learned classifiers to identify popular videos.

1. Introduction

User Generated Content (UGC) sites exemplified by YouTube are nowadays a major Internet phenomenon that entertains over a billion users—almost one-third of all people on the internet—and form a billion-dollar global industry [1]. Given the scale, the dynamics, and the decentralization of the contents provided by individual users, one fundamental question for maintaining and growing such UGC sites is to understand and to identify contents that will gain great popularities. The content popularity, implicitly measured by the number of views, has been extensively studied before, ranging from revealing the popularity characteristics [2–7] to popularity predictions based on features [8–12] and generative models [13,14]. In this paper, we revisit this problem and complement previous studies with two new aspects, as follows:

- *Implicit and explicit popularity*: Implicit popularity in terms of the number of views reflect user's attention but not necessarily appreciations. Nowadays, UGC sites provide many opportunities for users to interact with the contents. Can we infer the content popularity explicitly from these interactions and predict the contents that users really like?
- *The benefits of the built-in social features*: For user retention and attraction, UGC sites are coupled with and have been introducing new social features like *donation* and *chat replay*. Can we leverage these new social features to refine the popularity prediction problem and particularly to make predictions immediately upon publication?

To this end, we need to choose a UGC site with enhanced social features as our research vehicle and obtain preferably the complete view or a representative sample of the whole system. And the dataset should be multi-dimensional and contains not only the content (video) information but also the user information including their relationships and interactions. For our analysis, we have chosen BiliBili [15], a Youtube-like UGC site with enhanced social features, for the following two reasons:

First, beyond traditional UGC functions like video sharing/viewing, voting, commenting and channel subscription, BiliBili implements a number of social features, including *non-reciprocal user following, chat replay*, and *virtual money donation*. Features extracted from social relationships like following have been shown to be informative for popularity prediction [9–14]. However, these

[*] Corresponding author at: School of Computer, National University of Defense Technology (NUDT), China.

E-mail addresses: ljia@cau.edu.cn (A.L. Jia), shensiqi@nudt.edu.cn (S. Shen), dsli@nudt.edu.cn (D. Li), chenshengling@baidu.com (S. Chen).

https://doi.org/10.1016/j.comnet.2018.05.004
1389-1286/© 2018 Elsevier B.V. All rights reserved.

studies only focus on a small subset of videos and predictions can only be made after they get promoted in external social networks. In contrast, the built-in social features in BiliBili provide the possibility and the extra information for predicting the popularity, even immediately upon publication, of any video.

Secondly, unlike previous studies that are often carried out on sampled datasets [11,13,14,16–18], we were able to capture the entire repository of BiliBili (at the time of our crawling) with over 2 million videos and over 28 million users. Moreover, the information we obtained includes not only the repository characteristics such as the video duration and the user gender, but also user activities and interactions, for example, how users view and donate to the videos, when and who left what comments, and how users follow each other. This global view and fine-grained information avoid potential defects, e.g., under- and over- estimations of certain network properties, caused by sampling and sampling biases [19–21].

Our analysis of BiliBili mainly consists of three parts:

Implicit and explicit popularity. We first quantitively reveal the scale and the characteristics of BiliBili by examining its video repository that contains over 2 million videos. We analyze the implicit and the explicit popularity and we study the influence of the video type. Similar to previous studies, we find that both the implicit and the explicit video popularity is *highly skewed*, with a small number of videos collecting a large portion of the total popularity. Interestingly, we further find that popularity metrics that measure user's attention but not necessarily appreciations (i.e., the number of views and the number of replayed chat messages) are best fitted by Log-Normal distributions, whereas metrics that directly reflect user's appreciations (the amount of donated virtual money and the number of favorites) are best fitted by more skewed Power-Law distributions. Moreover, while users prefer to view videos shared from other sites, they are more generous in donating virtual money to the videos uploaded locally (highly likely to be user-generated).

Social features. Intuitively, social features provide complementary information for inferring both the explicit and the implicit popularity. Based on a dataset containing information on over 28 million users, we first examine the characteristics of three types of user activities and interactions, i.e., uploading, following, and commenting. We derive a number of interesting findings including that uploaders not only get more followers but are also more active in following others and that in general male users are more active while female users are more popular. Then, we propose two graphs that look not only at who a user is connected to, but also how those connected users are linked amongst each other. These graphs capture both the direct user relationships and the higher-order social structures and they provide valuable information for the popularity prediction problem.

Popularity prediction. Finally, applying our findings, we build feature-based predictors that can successfully predict popular videos, in terms of the number of views (implicit popularity) and the amount of virtual money users donated (explicit popularity). As it turns out, both the social features and the graphs we proposed are informative for the popularity prediction problem, e.g., on a balanced dataset where random guessing would yield an accuracy of 50%, our predictors achieve 86% even without knowing the early view patterns, and further improves the accuracy to 92% with only one-day observation.

We summarize our contributions as follows:

- We collect, use, and offer public access[1] to the dataset that contains the whole repository of BiliBili (at the time of the crawling), with detailed statistics for 2,858,844 videos and 28,962,041 users (Section 2).
- We provide a characterization on BiliBili. Our analysis includes (i) the repository scale, (ii) the statistical properties of the video popularity (Section 3), (iii) the uploading activity, (iv) the following activity, and (v) the commenting activity of the users (Section 4).
- We propose two graphs to analyze the user relationships, i.e., a *follow graph* that contains 10,749,726 users and a *comment graph* that contains 6,677,456 users (Section 5).
- We build machine-learned classifiers to identify with high accuracies the popular videos (Section 6).

2. Methodology and the BiliBili dataset

In this section, we first give a brief introduction on the basic operations of BiliBili. Then, we identify important and informative characteristics and metrics for the popularity prediction problem. Finally, we introduce our measurement methodology and the dataset used throughout this article, and we describe the scale of BiliBili.

2.1. An overview of BiliBili

BiliBili is a YouTube-like UGC site with enhanced social features. As in traditional UGC sites, users in BiliBili can consume and share videos, vote and leave comments to videos, and subscribe to channels (of a series of videos). In addition, BiliBili provide three (unique) social features:

(i) *Non-reciprocal user following* that allows users to follow each other, for social purposes or merely getting updates on videos that they are interested in.
(ii) *Chat replays*, named *danmu* in BiliBili, are comments flying over the screen on exactly the video time when they are left before by various users. Chat replays allow the later users to understand and to communicate with their ancestors. They provide immersive viewing experiences and are adopted by a number of popular UGC sites including Twitch.tv [22]. An example of a BiliBili video page with chat replays is shown in Fig. 1.
(iii) *Virtual money donations* are made to uploaders by users who appreciate their contribution. In BiliBili, the virtual money (named coin) is used in various circumstances, including upgrading user membership and exchanging for new emojis.

Terminology. As users in BiliBili can take various roles, to simplify our arguments, we define the following user types:

(i) *uploaders*, users that have uploaded at least one video,
(ii) *viewers*, users that have not uploaded any videos,
(iii) *commenters*, users that have left at least one danmu, and
(iv) *social users*, users that have followed or have been followed by at least one other user. Specifically, if a user A follows a user B, then user A is named the *follower* of user B and user B is named the *followee* of user A.

2.2. Characteristics and metrics

To characterize BiliBili videos and users, we identify the following three important aspects that make up the basic operations of BiliBili, which provide important knowledge for the popularity prediction problem and will be later discussed in detail in Sections 3, 4, and 5, respectively.

[1] https://sites.google.com/view/bilibilidataset.