Information Sciences 339 (2016) 310–331

Contents lists available at ScienceDirect

Information Sciences

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

Exploiting concept drift to predict popularity of social multimedia in microblogs

Cheng-Te Li[a,*], Man-Kwan Shan[b], Shih-Hong Jheng[b], Kuan-Ching Chou[b]

[a] Research Center for IT Innovation, Academia Sinica, Taipei 115, Taiwan
[b] Department of Computer Science, National Chengchi University, Taipei 106, Taiwan

ARTICLE INFO

Article history:
Received 10 December 2013
Revised 28 November 2015
Accepted 1 January 2016
Available online 7 January 2016

Keywords:
Popularity prediction
Social multimedia
Concept drift
Information diffusion
Microblog social network

ABSTRACT

Microblogging services such as Twitter and Plurk allow users to easily access and share different types of social multimedia (e.g. images and videos) in the cyber world. However, the massive amount of information available causes information overload, which prevents users from quickly accessing popular and important digital content. This paper studies the problem of predicting the popularity of social multimedia content embedded in short microblog messages. A property of social multimedia is that it can be continuously re-shared, thus its popularity may revive or evolve over time. We exploit the idea of concept drift to capture this property. We formulate the problem using a classification-based approach and propose to tackle two tasks, re-share classification and popularity score classification. Two categories of features are devised and extracted, including information diffusion and explicit multimedia meta information. We develop a concept drift-based popularity predictor by ensembling multiple trained classifiers from social multimedia instances in different time intervals. The key idea lies in dynamically determining the ensemble weights of classifiers. Experiments conducted on Plurk and Twitter datasets show the high accuracy of the popularity classification and the results on detecting popular social multimedia are promising.

1. Introduction

Microblogging services, such as Facebook, Twitter, and Plurk, are platforms that allow users to share quick, short messages with friends. Such information can quick spread in the cyber world. Various types of digital content can be easily embedded in microblog messages and be shared between users. *Social multimedia*, such as images and videos, is one of the most accessible and wide-spread digital content available in microblogs. With interactive functionalities on microblogging services, including endorse, re-share, comment, and rate, social multimedia can gain exposure and popularity as more users positively acknowledge it. These functionalities facilitate the communication between users and help users access novel information in a timely manner. However, these social functionalities can also lead to information overload. With large volumes of messages appearing on the personal pages of users in a short time period, users tend to overlook important messages. This information overload may prevent users from receiving popular and important social multimedia quickly, which

* Corresponding author. Tel.: +886-2-2787-2370.
E-mail addresses: ctli@citi.sinica.edu.tw, reliefli@gmail.com (C.-T. Li), mkshan@nccu.edu.tw (M.-K. Shan), shjheng@gmail.com (S.-H. Jheng), johnny1991524@gmail.com (K.-C. Chou).

http://dx.doi.org/10.1016/j.ins.2016.01.009
0020-0255/

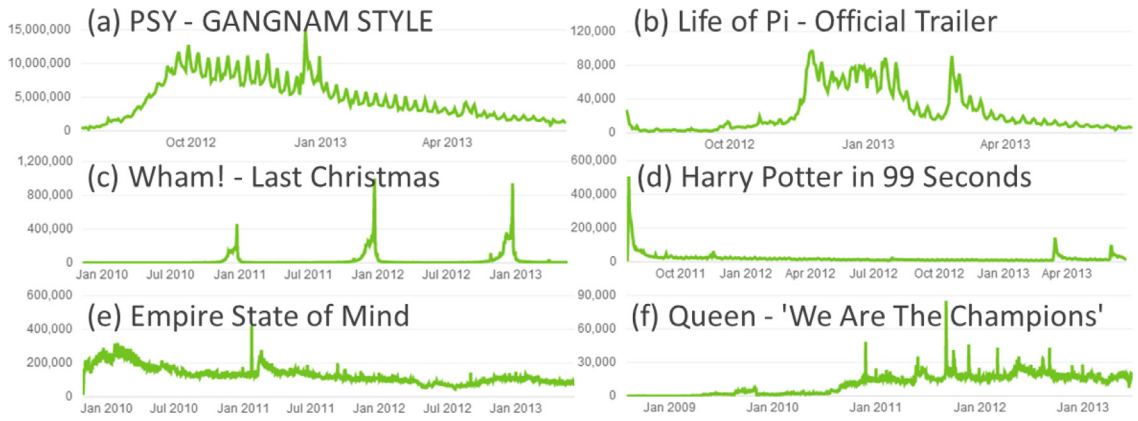

Fig. 1. Sequences of popularity for six YouTube videos.

especially affects users with lots of friends. Therefore, it is useful and appreciated to recommend popular social multimedia to such users so that they can quickly catch on to current trends and follow new users.

To discover popular social multimedia for recommendation, we propose to predict the popularity and spread of social multimedia embedded in a microblog social network. We resort to the following features of microblogs to formulate the problem and the solution. First, microblogs provide friendly user interfaces for real-time message posting. This real-time property enhances the visibility of information and provides the potential to quickly accumulate view counts for social multimedia embedded in messages. Second, the content of microblog posts tends to be conversation-based, with a sequence of responses. Conversations can boost the vitality of social multimedia through interactions between users. Third, users of a microblog service are connected by an underlying social network. Social multimedia embedded in messages posted by users can be viewed by their friends. Therefore, social multimedia can have an implicit or explicit impact on more users as they re-share such content. Last but not least, messages containing the same social multimedia can be regenerated and shared by users over time under different *uncertain* contexts. Each social multimedia content can (1) be instantly and widely spread in a very short time period (as shown in Fig. 1(b) and (d)), (2) remain stable in terms of popularity for a long period (as shown in Fig. 1(e) and (f)), (3) accumulate or dissipate its popularity gradually (as shown in Fig. 1(a) and (b)), and/or (4) alternate between being popular and unpopular (as shown in Fig. 1(c)). We will elaborate on the details of each scenario in the following sections.

The importance of understanding the popularity of online social multimedia is four-fold. First, for network and cloud service providers, accurate prediction of popularity at an early stage can help with planning sufficient storage and computation resources and reserving adequate bandwidth in advance to handle real-time streaming requests. The service providers can save resources while delivering smooth and high-quality multimedia contents. Second, for content producers, understanding currently trending topics is essential to developing content and diffusion strategies of new social multimedia. A successful social multimedia campaign attracts a large number of viewers, which directly affects profit, fame, and social influence. Third, for content consumers, they can receive accurate recommendations on trending multimedia items at early stages. Fourth, for advertisers, high popularity of social multimedia contents corresponds to high revenue. Finding popular multimedia contents in an accurate and prompt manner leads to more effective and efficient targeted marketing that improves profitability with lower costs.

Several studies have tried to predict the popularity of *messages* (i.e., *meme*-styled short texts) in microblogs. Diverse features have been explored, including LDA topical features and social network features [18], the effect of early adopters [3], the subjectivity of the message language [5], sentiment features [32], and the temporal patterns of the popularity evolution [1]. However, there exists a fundamental difference between predicting the popularity of messages and social multimedia. The life cycle of short messages on microblogs (e.g. retweeted messages in Twitter) usually lasts for 1–2 weeks [49] because the real-time property of short messaging leads to a huge volume of information, while novel events are being created daily. However, social multimedia (e.g. images and videos) embedded in microblog messages exhibit a longer life cycle since these digital contents are usually hosted on external media-sharing platforms (e.g. YouTube and Flickr), where media contents are allowed to be repeatedly searched and shared over time. Such fact implies that, instead of bursting onto the scene and cause an instant revolution, the popularity of social multimedia may be revived a long time after its initial publication, thus making popularity prediction of social multimedia challenging.

Our goal is to predict the popularity of social multimedia in a microblog social network. Images and YouTube videos embedded in messages are considered as the targeted social multimedia in this study. Given a social multimedia, which is embedded in a message posted in a certain time interval t, with all the relevant information before t, we aim to exploit the machine learning technique to predict the popularity of the targeted social multimedia at time interval $t+1$. To facilitate the prediction task, we define the popularity of a social multimedia in terms of time intervals. We treat the popularity prediction as a classification problem, which can be divided into two prediction tasks, *re-share classification* and *popularity*