Information Sciences 369 (2016) 585–613

Contents lists available at ScienceDirect

Information Sciences

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

Predicting the popularity growth of online content: Model and algorithm

Ilias N. Lymperopoulos

Department of Management Science and Technology, Athens University of Economics and Business, 47a Evelpidon Str., Athens, 11362, Greece

ARTICLE INFO

Article history:
Received 9 June 2015
Revised 25 June 2016
Accepted 17 July 2016
Available online 18 July 2016

Keywords:
Popularity prediction
Information diffusion
Information cascade
Social network
Online content adoption
Popularity growth model

ABSTRACT

The evolution of the popularity of online content is analyzed, and two characteristic patterns pertaining to linear and non-linear growth periods are detected. While the former characterizes the propagation of online content through a dynamical process in a state of statistical equilibrium, the latter appears when this state is perturbed by exogenous intervention events. Such episodes increase the susceptibility of higher threshold individuals who opportunistically adopt the propagating content. To capture the dynamics of both diffusion modes, the popularity of online content is modeled by interlacing linear and non-linear growth terms, reduced to 1st-degree polynomial and logistic functions corresponding respectively to stationary and non-stationary adoption phases. The precise fit of the model to empirical popularity patterns verifies its suitability as prediction tool. The proposed model is employed to generate forecasts about the popularity of online content through extrapolation. Highly accurate prediction results surpassing existing methods in terms of precision and predictive capacity are demonstrated. The prediction method is formulated into an algorithm, applicable to real time forecasting of the popularity of online content without training, using minimal, macroscopic, publicly available information.

1. Introduction

As online social networks become ubiquitous and develop into the primary venue for the dissemination of information generated by individuals, businesses and organizations, the prediction of the popularity of online content is increasingly important for reputation management, growth of business opportunities, and effective communication. Reliable forecasts provide valuable insights into the content quality and its potential reach, thereby either corroborating the online communication efficiency, or providing timely signs for remedial action. Undoubtedly, forecasting the diffusion of online content is vital to a successful online presence, but nonetheless it begs the question: Is it possible to make predictions when a multitude of unforeseen factors affect the adoption dynamics?

The diffusion of online content is a dynamical process exhibiting time-varying behavior, and therefore prediction is inherently interlinked with the temporal characteristics of the popularity evolution. To formulate a prediction method, we follow a three-stage approach comprising the detection of universal popularity growth patterns, their modeling as a function of time, and the use of the derived model for generating popularity forecasts. The fitting of the model to real popularity patterns is essential in order to provide sufficient evidence of its adequacy to describe the popularity growth process, thereby

E-mail address: eliasliber@aueb.gr, eliasliber@hotmail.com

http://dx.doi.org/10.1016/j.ins.2016.07.043
0020-0255/

justifying its use as a prediction tool. The estimation of the model parameter values from available information makes it possible to predict the evolution of the online content popularity through forward extrapolation.

The implementation of the foregoing approach starts with the analysis of empirical data. In this regard, we examined time-series describing the popularity growth of Twitter hashtags, Memetracker phrases, and Flickr photos with a view to detecting common patterns, descriptive of universal principles in the underlying adoption dynamics of online content. The time-series are sample realizations from an infinite number of instances that could be potentially generated by an online content diffusion process. However, by inspecting samples from different online domains we can infer the statistical properties of the online content adoption dynamics that are invariant among all possible realizations, thereby fulfilling a primary objective of statistical investigation.

The data analysis showed that the popularity of online content evolves through two operating modes. The first one pertains to the steady-state of a stochastic dynamical process, whereby the adoption rate follows a stationary pattern, indicative of the statistical equilibrium state [6,15] of a content diffusion process. In this state the probability laws driving the content propagation remain invariant, thereby rendering the average adoption rate constant and independent of time, though small fluctuations still appear in the diffusion pattern. During this state the popularity growth of online content follows a linear pattern. Such an operating mode emerges when an online social network functions as a closed system, where the content adoption is driven by the activity of a homogeneous population of users who search for information and share links with their neighbors. Such events are mostly independent and result in trivial adoption cascades causing minor fluctuations in a stable adoption rate characterizing a content diffusion process in a state of statistical equilibrium. Stationary adoption patterns represent an enduring operating state and dominant condition in the online content diffusion for considerable stretches of time, especially for persistent online content. The second mode in the online content adoption process results from perturbations caused by the contact between an online social network and its surroundings. This contact takes place through intervention events significantly affecting the content adoption pattern which acquires the form of the individuals' response to information flows, such as news, advertising, and high-profile events. The exogenous excitation changes the susceptibility of a subpopulation of users with heterogeneous activation thresholds, thereby generating conditions favoring the opportunistic adoption of the propagating online content. The analysis of empirical data showed that the popularity evolution patterns generated by intervention events constitute a single class described by the logistic growth dynamics [30,42]. This means that the adoption of online content during non-stationary diffusion periods is analogous to the adoption of innovations [5,34,40], threshold phenomena in collective behavior [12,17], and other self-limiting growth processes pertaining to biology, ecology, epidemics and demography [30]. Indicative examples of the popularity evolution of online content are depicted in Fig. 1. The universality of the illustrated patterns among all types of online content allows the generalization of the popularity growth dynamics, which can be expressed in the form of a model comprising intermingled linear and logistic growth phases much like an input-dependent renewal process consisting of hold and jump periods [9]. By modeling these patterns and estimating the values of the model parameters from real data, we demonstrate that we can accurately predict the popularity growth of online content regardless of the time resolution of the available information, the type, popularity level and persistence of content, as well as the length of the prediction horizon. The proposed prediction method is differentiated from existing approaches in many respects. In particular:

- It provides precise estimations of the popularity growth of online content in linear and non-linear adoption periods using minimal, macroscopic, real time information. To the best of our knowledge such forecasts are not generated by existing methods.
- It is based on a prediction model accurately fitting time-varying popularity patterns. Existing methods do not provide evidence of fitting dynamic adoption patterns.
- It is universal and applicable to the popularity prediction of any type of online content, irrespective of the content popularity and persistence.
- It can be applied both to the prediction of the number of adopters, and to the prediction of the activity level in relation to a particular piece of online content.
- It does not require training and information about the content attributes, the users' characteristics, the social network structure, and the early adoption patterns.
- The modeling of popularity growth patterns is independent of the time resolution of the available data, thus enabling the generation of accurate predictions using various timescales.

The simplicity, generality, flexibility and accuracy of the proposed method are its major strengths contributing to an effective solution to the online content popularity prediction problem in real time with limited data. The rest of the paper is organized as follows. In Section 2 we review studies relevant to the popularity prediction of online content. In Section 3 we utilize the phenomenological observations in order to construct an equation-based model fitting the popularity evolution of online content. In Section 4 we discuss how the proposed model can be incorporated into a method aiming at generating popularity forecasts using information about the adoption rate and its acceleration. In Section 5 we provide a series of experiments confirming: (a) The effectiveness of the proposed method; (b) the substantial improvement on the performance of benchmark models; and (c) the highly precise fitting of the equation-based predictive model to real popularity patterns. In Section 6, we provide an algorithmic implementation of the proposed method addressing the popularity prediction of online content in real time. Finally, in Section 7 we discuss the findings of this study and present the concluding remarks.