Electronic Commerce Research and Applications 14 (2015) 150–168

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

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

Predicting microblog users' lifetime activities – A user-based analysis

Xi Chen [a], Ruibin Geng [a], Shun Cai [b,*]

[a] School of Management, Zhejiang University, China
[b] School of Management, Xiamen University, 422 South Siming Road, Xiamen City, China

ARTICLE INFO

Article history:
Available online 12 June 2014

Keywords:
Social network
Microblogging
User lifetime vitality
Pareto/NBD model
BG/NBD model

ABSTRACT

With the rapid development of online social media, social networking services have become an important research area in recent years. In particular, microblogging as a new social media platform draws much attention from both researchers and practitioners. Although most current studies focus on the effect of social networks on the diffusion of services or information, most are descriptions or explanations of what has already happened. This study focuses on future activity by employing probability models such as the Pareto/NBD and BG/NBD models to predict user lifetime vitality. Three experiments were implemented to test the two models. Our results showed that both the Pareto/NBD model and the BG/NBD model were effective in predicting SNS user usage behavior on microblogging websites. It was found that tweeting behavior is more suitable for such probability models than retweeting behavior and user segmentation can improve prediction accuracy by distinguishing between currently active and inactive users.

1. Introduction

With the rapid development of online social media, social networking services (SNSs) have become an important research area in recent years. We are now in the era of social colonization (Cortizo et al., 2011), in which technologies such as Facebook Connect and Google Friend Connect have standardized social functionalities among a vast majority of websites. Social networking services, including microblogs, have entered people's daily life and become an important social platform for their daily communication. Twitter—as the most famous and the earliest application of microblogging technology—already had 645 million registered users around the world as of January 2014 (Twitter Statistics, 2014). Microblogging is a broadcasting platform based on the follow mechanism in social networks, where users publish (tweet) and distribute (retweet) brief messages (usually of less than 140 characters) through Web or Wap client components. In August 2009, Sina launched its microblog called Sina Weibo, one of the first microblogging services in China. Just two years later, it became one of the most popular websites in China and a significant communication platform. The rapidly expanding trend of microblogging suggests that there could be significant business opportunities to explore. For instance, the microblogging platform helps build online word-of-mouth (WoM) and facilitates social shopping (Liang et al., 2012, Amblee and Bui, 2012, Olbrich and Holsing, 2012). Although microblogging services have attracted a huge number of users, however, nearly half of registered users are inactive "zombie" accounts (Fu and Chau, 2013). How to evaluate the vitality of users on an SNS platform has become one of the primary concerns of platform operators.

Although huge numbers of users have registered on various SNSs, only active users are the major contributors to the platforms. The proportion of active users in the user base is an important measure of the healthy development of SNS platforms (Valenzuela et al., 2008). Predicting future vitality of SNS users has great significance for service providers, especially online platform operators. Sina Weibo, for example, treats user vitality analysis as direct guidance for the operation and management practices of microblogging products (such as micro-applications, and micro-topics) (Annual report by Sina.com 2013). Specifically, operators can track the interaction among users who are involved in a certain micro-topic or activity and predict the future popularity trend of the topic or activity by evaluating the current vitality of the user base. Users on the platform are assessed as having different vitality levels according to how often they log in and post messages. The lifetime vitality of registered users relates to the future development of the SNS platform, simultaneously affecting the confidence and enthusiasm of other enterprises in microblogging marketing and advertising.

Besides having managerial importance, user lifetime vitality predictions are also of great theoretical importance. Existing studies on social networking services largely consist of static descriptions or explanations of what has already happened. They usually focus on understanding user motivation and intention to use a certain IT artifact (Java et al., 2007, Marios, 2002). Another stream of

* Corresponding author. Tel.: +86 592 2187051.
E-mail addresses: chen_xi@zju.edu.cn (X. Chen), grace.bin1207@gmail.com (R. Geng), caishun@xmu.edu.cn (S. Cai).

http://dx.doi.org/10.1016/j.elerap.2014.06.001
1567-4223/

research focuses on user-generated content (UGC) with social networking behavior and networking features as explanatory variables (e.g. Susarla et al., 2012). Sentiment analysis and opinion mining are applied on UGC for the sake of trend prediction and anomaly detection. However, such static analysis and content-based research on social networking services has several limitations:

- *Static analysis*. The different frameworks and indicators to evaluate user activity and influential power that have been proposed in the literature are mainly based on historical behavior data or the static network formed by behavior history. Static analysis largely neglects the dynamic nature of fast-moving social network interaction.
- *UCG-based analysis*. Billions of SNS users generate numerous words and sentences every day. Sentiment analysis and text mining are meaningless if they are not concentrated on specific topics or key words. Microblogging as the new representative social networking service places more emphasis on interaction between users, which is quite different from the content-based network services of Web 2.0, such as YouTube.

To complement existing literature on SNSs, rather than explaining users' *past* behavior, this research focuses on predicting users' *future* behavior. In particular, this research employs stochastic models to predict the lifetime vitality of individual microblogging users' behavior. Empirical studies are conducted using microblog user interaction data from Sina Weibo. The customer base analysis models (i.e. Pareto/NBD model and BG/NBD model) are rooted in customer lifetime value theory in marketing research (Jain and Singh, 2002), which makes use of stochastic characteristics of interaction behavior to predict future behavior. Applying customer base analysis models to the prediction of SNS user behavior to gain new insights would contribute to current literature in several ways. First, a customer base analysis model goes beyond historical description and aims at predicting the future value and the continuous vitality of the customer base. Second, a customer base analysis perspective focuses on user behavior rather than on generated content. Analysis and prediction of platform vitality is based on an ensemble of individual behavior characteristics. This perspective distinguishes new social media from traditional websites. The unit of analysis in this paper is individual behavior rather than the microblogs created by a user. Finally, by focusing on behavior rather than only on generated content as an analysis unit, it becomes possible to further segment different types of social networking interaction. In general, users may tweet original words or retweet others' tweets; these two types of behavior may have different characteristics and potential for predicting future vitality.

This paper is organized as follows. Section 1 introduces our research. Section 2 outlines related work on microblogging user behavior prediction and customer base analysis models. Section 3 describes the research method composed of parameter estimation process and a derivation of two key expressions for managerial insights. Section 4 illustrates empirical test results of the two models based on user interaction data from Sina Weibo. Section 5 discusses the empirical results. Finally, Section 6 draws conclusions and suggests future research directions.

2. Literature review

2.1. Microblogging user behavior prediction

Recently, microblogging user behavior has attracted much attention from researchers and practitioners. A look at prior literature reveals that a large proportion of it attempts to address the following four research areas: (i) how users access microblogging services, (ii) how users generate abundant content, (iii) how users establish social relationships, and (iv) how users interact with others in a network. Scholars discovered how users access microblogging services by analyzing the sequence of web pages viewed or the path taken by users when they navigated a microblogging website. For example, Teevan et al. (2011) revealed the way in which people search socially generated content to differentiate it from traditional websites using large-scale query logs. User-generated content is one of the most important outcomes of social networking services. Common methods applied in studies based on user-generated content are text mining and semantic analysis. Weerkamp and de Rijke (2012) conducted representative research in which they explored a set of activities that were likely to become popular at a later time using time-aware information extraction of users' tweets. Understanding how users establish social relationships in microblogging networks is important for many tasks such as friend recommendation and community detection. Communication intention prediction by Chelmis and Prasanna (2012) and link formation prediction by Yin et al. (2011) both contributed to this research question. In addition to behaviors discussed above, users' interaction behavior is equally crucial to sustainable development of a microblogging platform. Our study focused specifically on interaction behavior prediction.

Users interacted with each other in a microblogging network by posting messages or reposting others' microblogs. Information is diffused through such behavior on social networks. Thus existing literature related to interaction behavior mainly focuses on the information diffusion mechanism. Information diffusion prediction reveals the characteristics of the news or ideas spreading through social networks, which is one of the key topics discussed in online social network research (Liben-Nowell and Kleinberg, 2008, Chen et al., 2013). Machine learning algorithms (such as collaborative filtering) and Markov random process are commonly used approaches to predict the probability of a tweet being retweeted (Zaman et al., 2010, Song et al., 2007, Artzi et al., 2012). Zhao et al. (2013) proposed a hybrid linear regression to predict the occurrence count of given hashtags within a specific time window. The model studied dependencies between various feature types, including hashtag content, global tweet features, graph topology features, and global temporal features, with the spread of information. Song et al. (2012) established the likelihood of impacts on user behavior of four key factors: user authority, user activity, user preferences and user social relations from an empirical analysis perspective. A support vector machine (SVM) was used to classify user behavior patterns (whether a message would be forwarded or ignored by users) and to predict the information diffusion mechanism. However, user behavior patterns are much more complex than a traditional "yes" or "no" binary. Although the aim of these papers is behavior prediction among microblogging users, the unit of analysis is still the message or content created by a user. In this paper, we focus on individuals who tweet or retweet on a microblogging platform rather than on the microblogs that are tweeted or retweeted.

Microblogging users are customers of social networking services platforms in a broad sense. Although most rarely contribute monetary value to the SNS platform directly, they spend much time and effort posting messages and interacting with friends on the platform and form the foundation of the SNS. To maintain and increase the number of active users in an SNS, microblogging platform operators needs to understand the current status of the user base (how many active users there are on the social media platform) and make reasonable predictions of users' future vitality (how active users will be in future). Aiming at providing a tool to better understand the SNS user base, we investigated the temporal dynamics of user microblogging behavior by applying customer base analysis.