Microblog Users’ Life Time Activity Prediction

Ruibin Geng  
*Zhejiang University, grace.bin1207@gmail.com*

Xi Chen  
*Zhejiang University, chen_xi@zju.edu.cn*

Shun Cai  
*Xiamen University, pchau@business.hku.hk*

Follow this and additional works at: [http://aisel.aisnet.org/pacis2013](http://aisel.aisnet.org/pacis2013)

**Recommended Citation**

[http://aisel.aisnet.org/pacis2013/18](http://aisel.aisnet.org/pacis2013/18)

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2013 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
MICROBLOG USERS’ LIFE TIME ACTIVITY PREDICTION

Ruibin Geng, School of Management, Zhejiang University, Hangzhou, China, 
grace.bin1207@gmail.com
Xi Chen, School of Management, Zhejiang University, Hangzhou, China, 
chen_xi@zju.edu.cn
Shun Cai, School of Management, Xiamen University, Xiamen, China, 
pchau@business.hku.hk

Abstract

With the fast development of online social media, social network services have become an important research area nowadays. We are now in the era of social colonization, in which technologies such as Facebook Connect and Google Friend Connect have standardized social functionalities among a vast majority of websites. Particularly, microblog as a new star needs more attention. Although most of current studies have focused on the effect of social network on the diffusion of services or information, usually those studies are descriptions or explanations of what already has happened. Limited study has been conducted focusing on SNS users and analysing their behaviours dynamically. In this paper, we used probability models such as Pareto/NBD and BG/NBD to predict customer lifetime vitality. The data we used include information on users’ tweet and retweet behaviour, such as recency and frequency. Our results showed that both Pareto/NBD model and BG/NBD model showed effective ability to fit and predict SNS users’ usage behaviour on microblog website. Tweet behaviors are more suitable for such probability models than retweet behaviors. Managerial implications of the two models should be highlighted as well. Interaction rate and dropout rate can be considered as the vitality index of the whole user base measuring how active users are and how likely a user is active. Managerial questions such as how active the users are in this platform now and how active the users will be in the future can be answered by applying those models.

Keywords: Social Network, microblogging, customer lifetime vitality, Pareto/NBD model, BG/NBD model.
1 INTRODUCTION

Nowadays with the rapid developing pace of social media network service based on Web 2.0 background, we have rushed into 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. As the expansion of online social network services, especially microblogging, social media have great impact on user behaviour and social development.

Twitter, as the most famous and the earliest application of microblog, already has 75 million registered users around the world since January 2010. As it is defined, microblogging is a broadcast platform based on the follow mechanism in social network, where users publish (tweet) and distribute (retweet) brief messages (usually less than 150 words) through Web or Wap client components. In August 2009, Sina launched Sina Weibo (Microblog) as a beta version which is the first microblogging service provided in China. Merely two years (in 2011) after Microblogging officially entered the Chinese Internet, it is already one of the most popular websites in China. The expanding trend of microblogging registered users scale in China indicates that there are great business opportunities and customer values worth of digging.

Users’ interaction information is regarded as one of the most valuable assets to Social Network Service (such as Microblogging) operation platform. Predicting future vitality of social network site has great significance to service providers, especially the online platform operators. Sina Weibo, for example, attached importance to user vitality analysis as direct guidance to operation and management practice of microblogging products (such as micro-activities, and micro-topics). To be specific, operators could track on the interaction behaviours of users who are involved in a certain micro-topic or activity and want to evaluate current vitality of the user base and predict the future popularity trend of the topic or activity. The lifetime vitality of registered users is straightforward related to the future prospects for the development of SNS platform, at the same time, affecting confidence and enthusiasm of other enterprise in microblogging marketing and advertising.

Traditional behavioural researches on user online activities (Marios 2002) mainly use complicated psychological theories, such as TAM (David 1989). Those researches usually focused on users’ motivation and intentions of using a certain IT artefact. However, existing academic researches and industry investigations on social network services have several limitations in business application.

- Static analysis. Different frameworks and indicators to evaluate user activity and influential power have been proposed by institutions and scholars. Such frameworks or indicators are mainly based on the historical behaviors data or the static network formed by history behaviors. The static analysis neglects the dynamic trait of the fast moving social network interaction.
- User generated content based. Another branch of research interests is focused on user generated content (UGC) with social network behaviors and networking features as explanatory variables (Susarla et al. 2011). Sentiment analysis and opinion mining are applied on UGC for the sake of trend prediction and anomaly detection. Microblogging as the new representative social network service puts more emphasis on the interaction between users, which is quite different from content-based network service of Web 2.0, such as YouTube.

In this paper, we built stochastic models to predict lifetime vitalities of microblogging users’ behaviour based on customer base analysis models. We conducted empirical analysis using microblog users’ interaction data from Sina. The customer base analysis models we relied on has their root in customer lifetime value evaluation which makes use of stochastic characteristics of interaction behaviours and shows several advantages for Microblogging users’ interaction behaviour analysis. Firstly, customer lifetime value model aims to predict the future value and the continuous vitality of the customer base. It goes far beyond the historical descriptions. Secondly, customer lifetime value perspective focuses on user behaviour rather than the generated content. Analysis and prediction of the platform vitality is based on the ensemble of individual behaviour characteristics. This perspective
distinguishes the new social media from traditional networks. The unit of analysis in this paper is individual behaviours not the microblogs created by user. Thirdly, due to the behaviour as analysis unit, we can further segment different types of individual behaviour during the social network interaction. Generally as twitter has named, user may tweet original words or retweet others’ twitter. These two types of behaviour may have different characteristics and abilities to predict future vitalities.

In this paper, we focus on SNS users’ behavioural characteristics in consuming services provided by Sina Weibo. Customer base analysis models (i.e. Pareto/NBD model, BG/NBD model) are applied in SNS user interaction analysis and prediction. The main contributions of these analytic models are to evaluate the expected interaction frequency for the whole user base over time periods and to predict the expected number of interactions in the future. Several conclusions and discussions are made based on the empirical study results. The reminder of this paper is organized as follows. Section 2 outlines related work on customer lifetime value and customer base analysis models. Section 3 describes the research method composed of parameter estimation process and derivation of two key expressions for managerial insights. Section 4 illustrates empirical test results of the two models on user interaction behaviour data from Sina Weibo. Section 5 makes discussion about the empirical results. Section 6 draws conclusions and future directions.

2 RELATED WORK

Kotler (1996) defined valuable customer as “individual, household, or organization who brings more than certain amount of long-term benefits beyond the cost used to attract, sell or serve the customer”. The excess amount of benefits attributed to relationships with the customer during the entire lifetime is noted as “Customer Lifetime Value” (CLV). Customer lifetime value (CLV) is typically used to identify profitable customers and to develop strategies to target customers. Considerable amount of literature specified the quantification of CLV in various marketing situations and expanded the range of application from sales to services.

Jain and Singh (2002) summarized the literature in CLV research in mainstream marketing into three directions: calculation of CLV, customer base analysis, and normative models of CLV. Calculation of CLV aimed to formulate the calculation function through revenue and cost analysis, while customer base analysis concentrated to predict the probabilistic value of future customer transactions based on existing customer base information. In contrast with the first two directions’ aim of CLV quantification, the third direction mainly focused on managerial implications of CLV analytical models, such as using CLV to support pricing decisions (Blattberg & Thomas 1997). Given the empirical background of Microblogging service, it is difficult to specify the revenue or cost related with a certain user. But the future transactions prediction models in customer base analysis can be appropriately migrated to social network service consuming cases. User interaction behaviours can be adopted by equivalent to customer transactions.

RFM model is a general paradigm to describe the statistic feature of customer transaction behaviours in customer base analysis (Bult & Wansbeek 1995; Fader et al. 2005a; Kahan 1998; Liu & Shih 2005a, 2005b). The term is defined by Bult and Wansbeek (1995), in which R denotes recency that means the last time customer purchase a product or service during an observed time period, F denotes frequency that calculates the number of purchase a customer made during an observed time period, and M denotes monetary that represents amount money customer spend during an observed period. Hughes (1994) proposed a method for RFM scoring that involved using RFM data concerning to sort individuals into five customer groups. Stone (1995) suggested that different weights should be assigned to RFM variables depending on the characteristics of the industry. Fader et al. (2005a) linked the RFM paradigm with customer lifetime value by a stochastic mode based on the Pareto/NBD framework (Schmittlein et al. 1987) which is the fundamental model of customer base analysis in
marketing. Pareto/NBD model is derived by combining NBD model\(^1\) that detects customer purchase process with the second type of Pareto distribution\(^2\) that illustrates customer death curve. Fader et al. (2005b) simplified assumptions about exponentially distributed customer lifetime duration in Pareto/NBD model and developed BG/NBD model. Jerath et al. (2007) proposed a compromise between “continuous-time death” Pareto/NBD (dropout can occur at any point of time) and “transactional discrete death” BG/NBD (dropout occurs after each purchase with certain probability) by calculating customer dropout probability with periodic death opportunity (PDO) approach. Abe (2009) developed the hierarchical Bayes extension to Pareto/NBD model by replacing independent gamma distributions of purchase rate and dropout rate with a correlated multivariate lognormal distribution. Table 1 compares the different assumptions made in these customer base analysis models. To sum up, models classified as customer base analysis always try to estimate the probability distribution function which demonstrates the individual stochastic feature and heterogeneity across customers in transaction behaviour.

<table>
<thead>
<tr>
<th>Customer Analysis Model</th>
<th>Assumptions for individual</th>
<th>Assumptions of heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pareto/NBD (Schmittlein et al. 1987)</td>
<td>• Poisson purchase: active customers’ purchase behaviour follows Poisson process with rate ( \lambda ). ( P[X = x</td>
<td>\lambda, t] = \frac{(\lambda t)^x e^{-\lambda t}}{x!}, x = 0, 1, 2, \ldots )</td>
</tr>
<tr>
<td>BG/NBD (Fader et al. 2005b)</td>
<td>• Exponential lifetime: customer lifetime duration is distributed exponential with dropout rate ( \mu ). ( f(t</td>
<td>\mu) = \mu e^{-\mu t}, t &gt; 0 )</td>
</tr>
<tr>
<td>PDO/NBD (Jerath et al. 2007)</td>
<td>• Poisson purchase: identical with Pareto/NBD • Bernoulli processed lifetime: customer becomes inactive with probability ( p ) after each transaction. ( P(\text{inactive after jth transaction}) = p(1-p)^{j-1} ) ( j=1,2,\ldots )</td>
<td>• Heterogeneity in transaction rate: identical with Pareto/NBD • Heterogeneity in dropout rate: individuals’ dropout rate distributed according to beta distribution ( f(p</td>
</tr>
<tr>
<td>Hierarchical Bayes Extension (Abe 2009)</td>
<td>• Poisson purchase: identical with Pareto/NBD • Exponential lifetime: identical with Pareto/NBD</td>
<td>• Heterogeneity in transaction rate: identical with Pareto/NBD • Heterogeneity in dropout rate: individuals’ dropout rate distributed according to beta distribution ( f(\theta</td>
</tr>
</tbody>
</table>

\(^1\) NBD model: negative binomial distribution, refers to the discrete probability distribution of the number of successes occurs in a specific number of times before a failure in the Bernoulli experiment sequence.

\(^2\) Pareto distribution: a power law probability distribution, satisfying \( p(X > x) = (x / x_{\text{min}})^{-\alpha}, \) with pdf as \( f(x) = k x_{\text{min}}^{\alpha} / x^{\alpha+1}, x > x_{\text{min}}. \)
Besides the Pareto/NBD based frameworks in customer base analysis, related work has been done by modelling customer relationships as Markov Chains (Pfeifer & Carraway 2000; Oded et al. 2008). Markov Chain models estimate transition probabilities among certain levels of possible customer relationship states which are evaluated and segmented by customers’ economic value (i.e. Monetary in RFM). The expected net present value of customer relationship for an infinite horizon is calculated by the reward vector that evaluates the economic values of different customer relationship states and one-step transition matrix that is estimated based on frequency and recency data of customer purchase behaviour (Pfeifer & Carraway, 2000).

Etzion et al. (2005) firstly introduced customer lifetime value theory into online service research field. They calculated the customer lifetime value of e-commerce based on Markov Chain modelling, which brings a new insight for the application of customer lifetime value theory. Besides, Joo et al. (2011) proposed that adding social network properties (such as number of ties in relation network and tie strength) to RFM-based customer value model improves the explanatory power of the baseline model on users’ gift-giving behaviour provided by SNS platform. However, these applications of CLV models in the background of e-commerce or social network service still centred on customer purchasing behaviours linked with monetary measurement, which makes little breakthrough in the nature of value calculation. In the general situation of online social networking service consuming, as operation platforms do not directly charge users, there is no feasible way to measure users’ monetary values. So frequency and recency are the main factors in RFM paradigm to adopt customer base analysis in SNS background.

It is the original creation of this paper to apply customer base analysis models to SNS users’ interaction behaviour prediction and to evaluate SNS users’ lifetime vitality. We conducted empirical studies using Pareto/NBD model and BG/NBD model with user interaction data from Sina Weibo. We compared the results of the two models on tweet and retweet behaviour data and made further discussion about the managerial implications.

3 RESEARCH METHODS

There are several gaps we have to fill before applying the customer base analysis framework. Firstly, the monetary value of users’ interaction behaviours in RFM paradigm is difficult to assess. So we subtly select Pareto/NBD model and BG/NBD model which merely require the frequency and recency information of behaviours while M is not necessary. Secondly, customer purchase behaviours are usually low-frequency so that previous studies just divided the long period data into two parts, one for model training and the other for prediction testing. However, as to user interaction behaviours, high frequency and periodic regularity are the obvious distinctions from purchasing behaviours. We set week-length observation time window to separate the data into sequence and conduct model estimation for each interval in consideration of these particular features. Thirdly, there are different types of SNS interaction behaviours in contrast with purchasing behaviours. We assumed that different types of interaction behaviours follow different stochastic process, so tweet and retweet are considered independently in Pareto/NBD and BG/NBD modelling. Comment as another very common behaviour involves bilateral interaction, and the opportunity for comment is directly induced by other users’ feedback, rather than a stochastic process. So we only apply customer base analysis to tweet and retweet, regardless of comment. Lastly, according to Schmittlein et al. (1987), “opportunities for transactions are continuous and unobserved” and “time at which customer become inactive is unobserved or difficult to observe” are two general conditions of Pareto/NBD analysis. Basically SNS interaction behaviours meet these requirements but need more accurate definition on user “death”.

To predict microblog users’ life time vitality, it involves three sub-tasks. First is feature extraction from the interactive behaviours of SNS users. As it is argued above, recency and frequency in RFM paradigm are required as the inputs of user base analysis models. Second is stochastic feature analysis of SNS interaction behaviour. Similar assumptions are made that user interaction behaviour follows Poisson random process with heterogeneity in gamma distribution of transaction rates both in
Pareto/NBD and BG/NBD model, while customer lifetime duration is assumed as exponential distribution with gamma distributed death rate in Pareto/NBD model and Bernoulli process with Beta distributed death rate in BG/NBD model. Based on such assumptions, we will conduct empirical test on the two models including parameter estimation and calculation of two managerial expressions, thirdly. The basic idea for such models is to extract $F$ (note as $X=x$) and $R$ (note as $t_e$) values of users’ interaction behaviours within week-length time intervals, then to conduct parameter estimation with the historical data based on model assumptions, finally to calculate the two managerial expressions of user base vitality evaluation and individual future vitality prediction. The specific models are illustrated blow:

- **Model assumptions.** SNS users’ interaction behaviors are assumed to follow Poisson process (formula (1) in Table 1) with gamma distributed (formula (2) in Table 1) interaction rate $\lambda$. As to users’ lifetime duration in social network, user’s death (or dropout) is defined as long-period of inactive states (i.e. zero interaction in social network). Special cases exist when someone are recognized dead for a long period of time but abruptly revive to interact with his friends in SNS. In that situation, we will regard the revived person as a new user and the behavior data is recollected since the first interaction after his revival. In Pareto/NBD model, user lifetime duration follows exponential distribution (formula (3) in Table 1) with gamma distributed (formula (4) in Table 1) dropout rate $\mu$. In addition, BG/NBD model assumed that user lifetime duration follows Bernoulli process (formula (5) in Table 1) with Beta distributed (formula (6) in Table 1) dropout rate $p$.

- **Parameter estimation.** Both Pareto/NBD and BG/NBD model parameters are estimated using maximum likelihood estimation (MLE) method. For a random selected user, its interaction rate and dropout rate are unknown and independent, so according to Bayes’ theorem and marginalization rule, the likelihood function is computed by taking expectation of individual-level results over the integral mix distributions for interaction rate and dropout rate. The likelihood functions for random selected user of Pareto/NBD and BG/NBD are formula (7) and formula (8). The MLE method is to maximize the value of expression (9). As we can see, the likelihood function of Pareto/NBD model is quite complicated due to the evaluation of Gaussian hypergeometric function\(^3\) (note as $\sum_{i=1}^{n} F_i(\bullet)$).

$$L(r,\alpha,\beta,\beta|X=x,t_e,T) = \frac{\Gamma(r+x)\alpha'\beta'}{\Gamma(r)} \times \left\{ \frac{1}{(\alpha+T)^{\alpha'+\beta'}} + \frac{s}{r+s+x} A_\alpha \right\},$$

where for $\alpha \geq \beta$, $A_\alpha = \frac{\sum_{i=1}^{n} F_i(r+s+x,s+1; r+s+x+1; \beta-\alpha \alpha + T)}{(\alpha+T)^{\alpha'+\beta'}}$

and for $\alpha < \beta$, $A_\alpha = \frac{\sum_{i=1}^{n} F_i(r+s+x,s+1; r+s+x+1; \beta-\alpha \beta + T)}{(\beta+T)^{\alpha'+\beta'}}$  \(\text{(7)}\)

$$L(r,\alpha,a,b|X=x,t_e,T) = B(a,b+x) \Gamma(r+x)\alpha' \Gamma(r+a+b+x-1) + \delta_{x=0} B(a,\beta) \Gamma(r+a+b+x-1) \Gamma(r+x)\alpha'$$

$$LL(r,\alpha,\beta) = \sum_{i=1}^{n} \ln[L(r,\alpha,a,b|X=x,t_e,T)]$$  \(\text{(8)}\)

- **Core managerial expressions.** The great contribution of the two models is to answer two managerial questions. How active are the users in this platform now? And how active will the users be in the future? (Schmittlein et al. 1987) For the first question, $E(X)$, the expected number of transactions in a time interval, is entitled to the expected transaction volume for the whole user

\[^3\]Gaussian hypergeometric function can be obtained by numerical integration, $F(a,b,c;z) = \sum_{n=0}^{\infty} \frac{(a)_n(b)_n z^n}{(c)_n n!}$, when $n > 0$, $(x)_n$ is denoted as $x(x+1)\cdots(x+n-1)$, otherwise equals to 1.
base over time. E(X) is the estimation of aggregate activity integrated over the parameter distributions. It is more reliable to compute E(X) rather than simply pooling the frequency data to calculate the sample mean. To answer the second question, E(Y) is derived as the prediction of a particular customer’s future microblogging activity, given information about his past behavior and the parameter estimates. The expressions of E(X) and E(Y) in Pareto/NBD model are formula (10) and (11), while in BG/NBD model are formula (12) and (13).

\[
E(X(t)\mid r, \alpha, s, \beta) = \frac{r \beta}{(s-1)\alpha} \left[1 - \left(\frac{\beta}{\beta + t}\right)^{s-1}\right]
\]

(10)

\[
E(Y(t)\mid X = x, t_1, T, r, \alpha, s, \beta) = \frac{(r+x)(\beta+T)}{(\alpha+T)(s-1)} \left[1 - \left(\frac{\beta+T}{\beta + T + t}\right)^{s-1}\right]
\]

(11)

\[
E(X(t)\mid r, \alpha, a, b) = \frac{a+b-1}{a-1} \left[1 - \left(\frac{\alpha}{\alpha + t}\right)^{r+1} F(r + b; a + b - 1; \frac{t}{\alpha + t}) \right]
\]

(12)

\[
E(Y(t)\mid X = x, t_1, T, r, \alpha, a, b) = \frac{a+b+x-1}{a-1} \left[1 - \left(\frac{\alpha+T}{\alpha + T + t}\right)^{r+1} F(r + x, b + x; a + b + x - 1; \frac{t}{\alpha + T + t}) \right]
\]

(13)

Above all, BG/NBD model is a simplified alternative of Pareto/NBD model. The Gaussian hypergeometric function based on given parameters in BG/NBD can be approximately evaluated by polynomial sequence, which is easy to imply in EXCEL. But in Pareto/NBD model the Gaussian hypergeometric function is involved in likelihood function of parameter estimation, which increases the computing complexity for high quality prediction. In next section, we will conduct empirical test on these two models with data of user’s tweet and retweet behaviour in Sina Weibo.

4 EMPIRICAL STUDY: DATA AND RESULTS

Users’ interaction behaviours in Microblogging platform are tracked on a daily basis from October 1\textsuperscript{st}, 2011 to January 13\textsuperscript{th}, 2012, aggregately 15 weeks. The numbers of tweet and retweet microblogs are accumulated by week as the frequency input of Pareto/NBD and BG/NBD models. As the observing time interval is set as one week, the recency input ranges from 0 to 7. The time distance between the last microblog released time and the time interval start time as the recency input is calculated accurate to day. After extraction of frequency and recency values of users’ tweet and retweet behaviours, parameter estimation and model calculation of BG/NBD is implemented in EXCEL (Fader et al. 2005b) and Pareto/NBD model in MATLAB (Fader et al. 2005c).

4.1 Data Preprocessing

The original dataset donated by Sina is composed of the records of number of tweet and retweet microblogs by each user id in each day. The size of observation user base is at the magnitude of 100 thousand. We cumulated the number of tweet (or retweet) microblogs by each user id within one week (7 days) as the “frequency” (note as X) of tweet (or retweet). Besides, the last day with non-zero number of tweet (or retweet) microblogs by each user id of one week is used to calculate the “recency” (note as tx) of tweet (or retweet) approximately. For example, in the first week (from October 1st to October 7th, 2011), if the last record with non-zero number of tweet (or retweet) microblogs of user id “PBEHEGG” occurred at October 4th, the recency of tweet (or retweet) is 4 (the last non-zero date minus the first date of the week and plus one). One special case is that the user did not tweet (or retweet) any microblogs in this week. Then the frequency and recency are both zero. After extraction of frequency and recency from the raw data, we sampled one percent of users’ records from the observation data of each week (nearly one thousand tweet (or retweet) frequency and recency records) as training set to estimate the parameters of Pareto/NBD and BG/NBD models. The expected interaction frequency for the whole user base over time (E(X) in formula (10) and (12)) and the individual interaction behaviour prediction in future weeks (E(Y) in formula (11) and (13)) is...
computed on the entire dataset of each week. The empirical test results of Pareto/NBD and BG/NBD model are mainly three parts: parameter estimation, E(X) calculation, and E(Y) prediction.

4.2 Parameter Estimation

For the two models, maximum likelihood estimation method is applied using parameter estimating likelihood functions in formula (7) and (8). The parameter set \((r, \alpha, s, \beta)\) of Pareto/NBD model and \((r, \alpha, a, b)\) of BG/NBD model are estimated for each week (Figure 1 and 2). According to the distribution assumptions of interaction behaviour and lifetime duration in formula (2), (4), and (6), the expected interaction rate \(E(X)\) and the expected dropout rate \(E(Y)\) of the whole user base are derived as formula (14). During the observed 15 weeks, the expected tweet rates moved stably at the average level of 0.578 in Pareto/NBD model and 0.580 in BG/NBD model. The average level of expected retweet rates is 0.963 in Pareto/NBD model and 1.055 in BG/NBD model. Pareto/NBD model shows larger variation in estimating expected retweet rate than BG/NBD model.

As to the expected dropout rates of tweet and retweet behaviour, Pareto/NBD model shows drastic churning compared with BG/NBD model. We discovered different types of distribution shape for the parameters, which indicating the existence of heterogeneity in users’ behaviours and the dynamic nature of users’ behaviours as shown in Figure 3. It is discovered from our analysis that long-tail shaped pdf curve occurred most frequently both in gamma distribution and beta distribution. Bell-shaped pdf curve (left top in Figure 3) of beta distribution appeared in week 2, 11, 12, and 14 (tweet behaviour). Highly right-skewed curve of beta distribution occurred in week 10 (tweet behaviour) and week 14 (retweet behaviour), while similar shaped gamma distribution occurred in week 4 and 12 (retweet behaviour).

\[
E[X| r, \alpha] = r / \alpha, \quad E[Y| s, \beta] = s / \beta, \quad \text{and} \quad E[p| a, b] = \frac{a}{a+b}
\]  

(14)

To show the parameter estimation results, the expected tweet rate and retweet rate are depicted in Figure 1.

![Figure 1](image)

Parameter estimation results of interaction (tweet and retweet) rates in Pareto/NBD and BG/NBD models
Figure 2. Parameter estimation results of dropout (tweet and retweet) rates in Pareto/NBD and BG/NBD models

Figure 3. Probability density function curves of beta distribution and gamma distribution (horizontal axis is the value of random variable x, vertical axis is the pdf corresponding to certain x)
4.3 \( E(X) \) Calculation

Estimation of expected interaction frequency for the whole user base over time is one of the outputs of Pareto/NBD and BG/NBD model. The general idea in traditional statistical practice is to estimate the population mean by the average level of randomly selected sample. In Pareto/NBD and BG/NBD model, the expected interaction frequency of the whole user base is estimated based on parameter distribution integration (formula (10) and (12)) rather than simply pooling the sampled data for mean calculation. We sampled nearly one thousand records from the observation data of each week for parameter estimation and calculated \( E(X) \) given time period of 7 days as the approximation to population mean of the whole user base. The accuracy of \( E(X) \) estimation to actual mean of the dataset by Pareto/NBD and BG/NBD model is compared in Figure 4. The mean squared error (MSE) of the approximation by Pareto/NBD is lower than BG/NBD in tweet behaviour analysis, while in retweet behaviour BG/NBD estimation performed better (Figure 4).

Estimation of expected interaction frequency for the whole user base over distributions of interaction rate and dropout rate increases the robustness by avoiding accidental errors of simple random sampling approximation method. Another property of \( E(X) \) calculation enables the computation of interaction frequency over time, not only the fixed time period of observation. This advantage will play significant role in vitality monitoring of certain microblogging topic.

![Figure 4. Platform expected interaction frequency estimation by Pareto/NBD and BG/NBD compared with actual population mean](image)

4.4 \( E(Y) \) Prediction

Another type of important outputs made by Pareto/NBD and BG/NBD model is the future interaction vitality prediction based on individual historical behaviour. We substituted the estimated

<table>
<thead>
<tr>
<th>week</th>
<th>actual mean</th>
<th>Pareto/NBD</th>
<th>BG/NBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.0680</td>
<td>5.0121</td>
<td>4.0720</td>
</tr>
<tr>
<td>2</td>
<td>4.0450</td>
<td>4.2715</td>
<td>4.0004</td>
</tr>
<tr>
<td>3</td>
<td>4.1280</td>
<td>4.2197</td>
<td>4.3374</td>
</tr>
<tr>
<td>4</td>
<td>4.2370</td>
<td>4.1817</td>
<td>4.8477</td>
</tr>
<tr>
<td>5</td>
<td>4.0540</td>
<td>3.7199</td>
<td>3.5832</td>
</tr>
<tr>
<td>6</td>
<td>4.2790</td>
<td>4.2329</td>
<td>4.1011</td>
</tr>
<tr>
<td>7</td>
<td>4.0270</td>
<td>3.7834</td>
<td>3.7780</td>
</tr>
<tr>
<td>8</td>
<td>3.9330</td>
<td>3.8545</td>
<td>3.3155</td>
</tr>
<tr>
<td>9</td>
<td>3.9990</td>
<td>3.5226</td>
<td>3.9766</td>
</tr>
<tr>
<td>10</td>
<td>3.9460</td>
<td>3.4481</td>
<td>3.9565</td>
</tr>
<tr>
<td>11</td>
<td>3.6730</td>
<td>3.6389</td>
<td>3.6788</td>
</tr>
<tr>
<td>12</td>
<td>3.9130</td>
<td>3.7884</td>
<td>3.7197</td>
</tr>
<tr>
<td>13</td>
<td>4.0300</td>
<td>3.9543</td>
<td>3.8624</td>
</tr>
<tr>
<td>14</td>
<td>4.2470</td>
<td>3.7388</td>
<td>4.0839</td>
</tr>
<tr>
<td>15</td>
<td>3.9660</td>
<td>3.4308</td>
<td>3.4969</td>
</tr>
</tbody>
</table>

MSE 0.1046 0.1173

<table>
<thead>
<tr>
<th>week</th>
<th>actual mean</th>
<th>Pareto/NBD</th>
<th>BG/NBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.1480</td>
<td>6.1280</td>
<td>7.1955</td>
</tr>
<tr>
<td>2</td>
<td>7.6820</td>
<td>5.0220</td>
<td>7.3838</td>
</tr>
<tr>
<td>3</td>
<td>7.3860</td>
<td>7.7559</td>
<td>7.3715</td>
</tr>
<tr>
<td>4</td>
<td>7.4910</td>
<td>7.7559</td>
<td>5.5003</td>
</tr>
<tr>
<td>5</td>
<td>7.3620</td>
<td>7.2489</td>
<td>5.5575</td>
</tr>
<tr>
<td>6</td>
<td>7.2170</td>
<td>6.3258</td>
<td>6.0382</td>
</tr>
<tr>
<td>7</td>
<td>7.0740</td>
<td>6.5805</td>
<td>6.2970</td>
</tr>
<tr>
<td>8</td>
<td>6.7680</td>
<td>5.6326</td>
<td>6.7121</td>
</tr>
<tr>
<td>9</td>
<td>7.0040</td>
<td>6.9172</td>
<td>6.0110</td>
</tr>
<tr>
<td>10</td>
<td>7.2920</td>
<td>6.4380</td>
<td>6.2308</td>
</tr>
<tr>
<td>11</td>
<td>6.9430</td>
<td>6.0817</td>
<td>7.3730</td>
</tr>
<tr>
<td>12</td>
<td>6.7610</td>
<td>6.4088</td>
<td>6.2989</td>
</tr>
<tr>
<td>13</td>
<td>6.7540</td>
<td>6.7444</td>
<td>6.1371</td>
</tr>
<tr>
<td>14</td>
<td>6.4280</td>
<td>5.9954</td>
<td>6.9297</td>
</tr>
<tr>
<td>15</td>
<td>6.3270</td>
<td>5.8728</td>
<td>6.4805</td>
</tr>
</tbody>
</table>

MSE 0.8546 0.5859
parameters into formula (11), (13) and made tweet and retweet frequency prediction for the next week based on historical Recency and Frequency data of every user in the previous week. Thus we could take advantage of the continuous weekly data to evaluate the prediction performance of the two models. To assess the prediction capacity of the two proposed models, different measures are used, such as mean absolute deviation (MAD) and mean squared error (MSE). In tweet behaviour prediction, more than 10 million prediction results by Pareto/NBD and BG/NBD models are compared with actual next week frequency (Table 2). Overall mean squared error of Pareto/NBD prediction on tweet frequency is 38.053, which is a bit lower than that of BG/NBD. As to retweet behaviour prediction, the prediction performance of Pareto/NBD and BG/NBD models declined with mean squared error rising to 117.716 and 122.722, because retweet behaviour data contained a higher proportion of outliers. Figure 5 and 6 shows the evolution trends of actual weekly tweet and retweet frequency mean compared with prediction value mean of Pareto/NBD and BG/NBD models in each week. Both of the two models showed sufficient prediction abilities to user interaction behaviour with limited frequency and recency data.

In addition to the mean comparison of prediction value with actual weekly frequency, we sampled one thousand prediction records and made Pearson correlation analysis. The correlation of actual value and Pareto/NBD prediction is 0.762 (p-value=0.000), while the correlation of actual and BG/NBD prediction is 0.497 (p-value=0.000). Above all, we can conclude that Pareto/NBD model has better prediction performance.

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>SE</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>SE</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>tweet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>0</td>
<td>301</td>
<td>4.087</td>
<td>8.804</td>
<td>1000796</td>
<td>0</td>
<td>305</td>
<td>7.054</td>
<td>15.51</td>
<td>998725</td>
</tr>
<tr>
<td>tx</td>
<td>0</td>
<td>7</td>
<td>3.139</td>
<td>3.045</td>
<td>1000796</td>
<td>0</td>
<td>7</td>
<td>3.471</td>
<td>3.072</td>
<td>998725</td>
</tr>
<tr>
<td>T</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>1000796</td>
<td>T</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>actual</td>
<td>0</td>
<td>1000</td>
<td>4.059</td>
<td>9.114</td>
<td>1000796</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pareto/NBD</td>
<td>0</td>
<td>280</td>
<td>3.871</td>
<td>8.096</td>
<td>1000796</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BG/NBD</td>
<td>0</td>
<td>694</td>
<td>3.73</td>
<td>7.887</td>
<td>1000796</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAD of Pareto/NBD</td>
<td>2.499</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MAD of BG/NBD</td>
<td>2.499</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSE of Pareto/NBD</td>
<td>38.053</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MSE of BG/NBD</td>
<td>39.597</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Interaction frequency prediction by Pareto/NBD and BG/NBD compared with actual next week frequency

![E(Y) prediction of tweet](image1)

Figure 5. Pareto/NBD and BG/NBD prediction performance of tweet frequency compared with actual next week mean
Based on the results of empirical test, both Pareto/NBD model and BG/NBD model showed effective ability to fit and predict SNS interaction behaviour. Approximate Estimation of average vitality for the whole user base and lower mean square error of future vitality prediction indicate that Pareto/NBD has a better fitness to SNS users’ interaction behaviours. However, Pareto/NBD model is more sensitive to high-frequent data because of excess computation burden in parameter estimation. As the behaviour frequency increased to one thousand, the prediction values of Pareto/NBD model are invalid while BG/NBD model was still available. On the other hand, a portion of prediction power of BG/NBD model is sacrificed by using exponential lifetime assumption to geometric dropout process. Due to computation complexity, Pareto/NBD model failed to handle the cases with extra high interaction frequency, while BG/NBD showed the ability to represent the interaction behaviour of a broader range of users.

Customer based models achieved better performance for tweet prediction than retweet prediction, judging from the consistency of the two models on platform vitality estimation and the prediction accuracy of individually future interaction frequency. Two main reasons are proposed. Firstly, tweet behaviours remain at a lower level of interaction vitality than retweet behaviours both in aggregated and in individual. Customer base analysis models are more adapted to low-frequent data. Besides, the variation of retweet behaviours is 16.821, three times higher than tweet behaviours. Outlier values of frequency which are more than three times of standard deviation from the mean in retweet behaviour are double of that in tweet behaviours. Pareto/NBD failed when standardized value of tweet behaviour larger than 24 and standardized value of retweet behaviour larger than 10, while BG/NBD failed when standardized value of tweet behaviour larger than 140 and standardized value of retweet behaviour larger than 66. Secondly, tweet behaviours are individual intrinsic motivated, which results in stably fluctuated weekly frequency. Retweet behaviours are extrinsic influenced. Breaking news and active friend connects will lead to abrupt change in weekly frequency of retweet behaviour. Considering the properties of Poisson process, the probability of occurrence is the same for any two intervals of with equal length and the occurrence and non-occurrence in any interval is independent with any other interval. This assumption may be violated when special events happen which may influence the users’ intention to retweet behaviour. This problem may be addressed by performing segmentation of customer before parameter estimation and conditional expectation calculations.
In addition to the overall comparison between performance of Pareto/NBD and BG/NBD models on tweet and retweet behaviours, managerial diagnostics should be highlighted as well. Expected value of interaction rate and dropout rate can be considered as the vitality index of the whole user base measuring how active users are and how likely a user is active. Interaction rate $\lambda$ in Poisson process represents the average vitality while users are active regardless of the length of observation time period. Expectation of interaction rates computed by estimated parameter $r$, $\alpha$ varied through time series in Figure 1, which gives insights on abnormal fluctuation detection. Any unexpected ups and downs will alert operators to pay close attention. Further analysis on interaction rate fluctuation may associate environmental covariates, such as social relationship and social position, with user’s interaction rates by regression model (16). Similar regression model can be constructed for dropout rate $\mu$ or $p$. Three types of probability density function (pdf) curve of dropout rates are verified in beta distribution and gamma distribution with typical representative in Figure 3. Bell-shaped pdf of dropout rate indicates that the dropout rate is concentrated in a narrow value field. Right skewed and long-tailed pdf of dropout rate are more common which embody asymmetric characteristics of heterogeneity in lifetime duration.

$$\log(E(\lambda)) = \gamma_0 + \gamma_1 d_1 + \cdots \gamma_d d_d$$,  \(d_1 \cdots d_d\) are environmental covariates \(16\)

Two remarkable contributions of Pareto/NBD and BG/NBD models are made to managerial implications. $E(X)$ indicates how active are the users in the platform right now and $E(Y)$ responses to how active will the users be in the future. $E(X)$ as the estimation of expected interaction frequency for the whole user base can be used to define the boundary of high vitality users and low vitality users. $E(Y)$ prediction provides a baseline of frequency expectation on users’ historical behaviour data. If users’ vitalities turn out to decline from the expectation, we should carry out exact retention strategies to keep waverers active. Besides the prediction performance at individual level, aggregated forecasting based on whole user base will project the evolutionary trend of platform vitality (Figure 5 and 6), which brings noteworthy applications. For example, vitality recession of user base segments concentrating on different micro-topics or micro-activities can be detected by comparing the aggregated prediction of platform vitality with the average interaction frequency of the whole user base ($E(X)$).

6 CONCLUSION

In this study, we attempt to integrate theoretical model of customer lifetime value with user vitality analysis in microblog website. Pareto/NBD and BG/NBD models are applied to SNS users’ interaction behaviour analysis and prediction. Empirical test on users’ tweet and retweet microblogging data from Sina Weibo showed that the two models were effective for predicting SNS user’s interaction behaviours accurately. Our analysis identified several new directions for future studies. In this study we distinguished tweet and retweet behaviours in implementation of Pareto/NBD and BG/NBD models. It would be interesting to integrate those two different types behaviour when calculating the vitality of users of SNS. Furthermore, applying the models to micro-activities’ and micro-topics’ vitality monitoring and prediction will be attractive.

There is one limitation that the high frequency and wide range variation characteristics of users’ interaction behaviours will affect the long-term prediction capability of the models. The two models predicted ineffectively when users’ vitality changed significantly. Abrupt fluctuation in interaction frequency implies the migration of users’ vitality level, which may violate Poisson process assumptions. However, as to the whole user base, there always exist users at certain vitality levels regardless of individual changes. That is why the models show good fitness at aggregated level but lower prediction accuracy at individual level. One of the improvements could be done in future is to make coarse RFM-based segmentation of customers (Hughes 1994; Ha & Park 1998). Different segments have different estimated parameters representing different levels of vitality. An ensemble of models can be built and each fits to a different subset of data.
References


