Twitter user growth analysis based on diversities in posting activities

Structured Abstract

Purpose – Twitter reflects events and trends in users' real lives because many of them post tweets related to their experiences. Many studies have succeeded in detecting events along with real-life information from a large amount of tweets by assuming users as social sensors. To collect a large amount of tweets based on specific users for successful Twitter studies, we have to know the characteristics of users who are active over long periods of time. In This paper, we clarify the characteristics of growth users over a long time to strategically collect a large amount of specific users' tweets.

Design/methodology/approach – We explore the status of users who were active in 2012, and classify users into three statuses of Dead, Lock, and Alive. Based on the differences between the numbers of tweets in 2012 and 2016, we further classify alive users into three types of Eraser, Slumber, and Growth. We analyze the characteristic feature values observed in each user behavior and provide interesting findings with each status/type based on GMM clustering and point-wise mutual information.

Findings – From our sophisticated experimental evaluations, we found that active users more easily dropped out than inactive users, and users who engaged in reciprocal communications often became Growth type. Also, we found that active users and users who were not retweeted by other users often became Eraser type. Our proposed methods effectively predicted Growth/Eraser-type users compared with the logistic regression model. From these results, we clarified the effectiveness of five feature values per active hour to detect intended Twitter user growth for strategically collecting a large amount of tweets.

Originality/value – We focus on user growth prediction. To appropriately estimate users who have potential for growth, we collect a large amount of users and explore their status and growth after three years. Our research quantitatively clarifies the characteristics of growth users by clustering using robust feature values and provides interesting findings obtained by analysis. After that, we propose an effective prediction method for growth users and evaluate the effectiveness of our proposed method.

1. Introduction

Twitter, which is one of the Internet's most popular social media services, had 320 million active users per month at the end of December 2015 (Prosser 2016). On it, users post very short messages called tweets and share them with others. Two types of post are the reply, which sends messages to particular users, and the retweet, which cites other user tweets. By following other accounts, Twitter users can rapidly obtain beneficial information such as sports news and current trends and easily communicate with friends using replies. By using retweets, users can spread tweets posted by their favorite celebrities and musicians to their followers.

Twitter reflects events and trends in users' real lives because many users post tweets related to their experiences. By using such characteristics to help people in various life aspects, many studies have succeeded in detecting events such as earthquakes and influenza epidemics, along with real-life information from a large amount of tweets by assuming Twitter users as social sensors (Sakaki et al. 2010, Aramaki et al. 2011, Yamamoto and Satoh 2015). A sentiment analysis of tweets is known as an effective approach for obtaining people's reaction to products, services, and public opinions (Rajadesingan et al. 2015, Bollen et al. 2011, Canuto et al. 2016).

We believe that the collection of a large amount of tweets is important to achieve these objectives. A simple approach is to use the Twitter Rest API, which is published by Twitter officials and has various types of API. However, we cannot use the API more than 180 times per 15 minutes per API token. This means that we need to strategically collect tweets within the API limitation. Moreover, Edwards (2015) reported that inactive users who don't engage in posting activity are increasing as time progresses. Koh (2014) also reported that only 10.7% of users who registered in 2012 were still active in February 2014. Therefore, to collect a large amount of tweets based on specific users, we have to know the characteristics of users who are active over a long time. It has been suggested that such knowledge is also used as an effective orientation to collect data in other social network services such as Weibo and Tumblr.

In this paper, we clarify the characteristics of users who are active over a long time, using a large amount of Twitter users. Our dataset consists of approximately 3.3 million users across two periods: the users' feature values collection period from April 2012 and June 2013 and the users' status exploration period from June 2016. As robust feature values for user activity evaluation, we observe the average number of tweets, replies, and retweets per hour. We compare the differences among three statuses of Dead, Lock, and Alive and detect the feature values to classify these three statuses using Gaussian mixture model (GMM). Moreover, we classify alive users into three types of Eraser, Slumber, and Growth based on the difference between the numbers of tweets in June 2013 and June 2016. We analyze the characteristics of Growth-type users and propose a user growth prediction method to strategically collect a large amount of tweets.

The remainder of our paper is organized as follows. In Section 2, we discuss related works. In Section 3, we explain our collected dataset and describe its statistical feature values. In Section 4, we analyze the difference among three statuses and the characteristics of each status based on GMM clustering and PMI scores. In Section 5, we propose a prediction method for growing users and evaluate the effectiveness of our method, compared with a logistic regression model. We conclude the paper by briefly describing future works in Section 6.

2. Related works

This research is focused on the analyzing and predicting intended Twitter user growth based on feature values of user's behavior. Therefore, in this section, we summarize the related works in two parts: studies of Twitter user behavior analysis in Section 2.1 and user behavior analysis in other social network services in Section 2.2. Finally, we describe the position of our study in Section 2.3.

2.1 Behavior analysis on Twitter

Study of behavior analysis on Twitter and Twitter-like services is flourishing. Java et al. (2007) analyzed the characteristics of the Twitter network and discussed people's Twitter utilization. They clarified that people utilize it for communication and sending and collecting information, and Twitter communities are classified into several types. Cha et al. (2010) analyzed user features with influence by comparing the number of followers, followees, replies, and retweets and clarified the user features extracted by each evaluation metric. To effectively diffuse tweets, Wang et al. (2013) estimated not only user interests but diffusion capability. They recommended the optimal address to diffuse one's own tweets. Yamaguchi et al. (2014) analyzed transitions in posting activity on the basis of feature values such as the number of tweets, replies, and retweets per week. They automatically split users into several clusters, calculated the transition probability between clusters on the basis of sequences of cluster numbers, and clarified the characteristics of users whose activity level dropped on Twitter. Myers and Leskovec (2014) clarified the catalyst that increases a user's followers based on bursts of retweet diffusion. They analyzed follower networks with time-stamps and proposed a model for inferring new followers for each user. Yang and Counts (2010) compared blogs and Twitter from the viewpoint of their information structures. They concluded that users who tweeted less than 30 times a month have shorter tweet intervals than blog post-intervals, and a larger number of tweets denotes a smaller difference between the two intervals. Yamaguchi et al. (2015) assumed that a list name plays the role of a folksonomy tag for users included in each list, and they analyzed tagging networks by using lists on Twitter. Their analysis clarified that the number of bilaterally tagged user pairs is major in friend relationships despite the number of them being minor in Twitter. Mizunuma et al. (2014) focused on Twitter bursts, in which the number of tweets exploded compared with the average number of weekday tweets. They classified bursts into four classes based on duration and magnitude and detected relationships between Twitter bursts and real events. Yuan et al. (2016) analyzed the reciprocity of social interactions between user pairs and observed that best friends vary as time progresses. Based on such results, they proposed a model to predict the repliers and retweeters of a particular tweet considering friendship dynamics. Gong et al. (2015) split users into four types according to amount of activity on Twitter and classified their motivations into five types: information sharing, personal update, friend interaction, public interaction, and advertisement. They clarified that information sharing and personal update are the top two motivations of speaking out across all user types. Gurajala et al. (2016) clarified the characteristics of fake user accounts by comparing update time and day of the week between true and fake accounts. Chalmers et al. (2011) analyzed inter-tweet intervals and tweet frequencies for all non-replies and replies. They clarified that posting intervals are different between replies and non-replies. Ghosh et al. (2011) analyzed retweeting activity using two features, time-intervals and user entropy, and identified five retweeting categories: automatic/robotic activity, newsworthy information dissemination, advertising and promotion, campaigns, and parasitic advertisements.

2.2 Behavior analysis on other SNSs

Many studies have clarified various user behaviors and life-cycles on web communities and social network services. Danescu-Niculescu-Mizil et al. (2013) analyzed the linguistic changes of users of web communities by using a two-gram language model. They divided the life-cycles of users into two stages: a linguistically innovative learning phase in which users adopt the community's language and a conservative phase in which users stagnate. Dror et al. (2012) identified users who are about to quit in a question answering service and reported that the number of answers given by users and the number of best answers are important signals between users who are likely to quit and those who are not. Kawale et al. (2009) studied the problem of player churn in online role playing games and proposed a churn prediction model based on examining social influences among players and their personal engagements. Cheng et al. (2015) clarified the anti-social behavior characteristics of users in online discussion communities. They detected overly exacting posts as one type of anti-social behavior and identified users with such behaviors. To predict the reply time of each user, Navaroli and Smyth (2015) built a model for activity patterns on each day and at each time and estimated the effective response time when users could reply with high probability. Tang et al. (2015) proposed a method of predicting negative links for each user. To achieve such a task, they defined three types of features: network features such as indegree and outdegree and cluster coefficients of each user, content-oriented features such as number of articles with positive and negative opinions, and interactive features such as number of communication behaviors with positive and negative interactions between two users.

2.3 Position of our study

Compared with the above studies, we focus on user growth prediction. To appropriately estimate users who have potential for growth, we collect a large amount of users and explore their status and growth after three years. Our research quantitatively clarifies the characteristics of growth users by clustering using robust feature values and provides interesting findings obtained by analysis. After that, we propose an effective prediction method for growth users and evaluate the effectiveness of our proposed method.

3. Dataset

3.1 User collection

In this section, we explain the details of our dataset for future user status and growth analysis. We exhaustively collected tweets from April 1, 2012 to June 5, 2013 (430 days) using the Twitter Search API in Japanese and randomly extracted 3,352,319 users and their tweets. The frequency distribution of the number of users per number of tweets and per number of active hours, which were observation times of user behavior such as tweets, replies, and retweets in an hour, are shown in Figures 1 and 2, respectively. In common, the vertical and horizontal axes are scaled by a common logarithm. The number of users who posted only one tweet was approximately 10,000. In Figure 2, the number of users increased, from which the number of active hours is approximately 10,000 because the maximum observation number of active hours is 10,320 (=430×24) in our collection period.

Next, we explored the status of these users from June 20 to June 30, 2016 using the Twitter Rest API. There are three kinds of status: Alive status is a user whose account existence we confirmed; Lock status is a locked account although it is alive currently and Dead status is a user whose account existence we couldn't confirm. The total number and ratio of users in each status are shown in Table 1. We can confirm that 64.5% of users are still active after three years and that 35.5% of user accounts cannot be accessed for collecting tweets.

For user growth analysis, we extracted alive users who created an account in the period between April 1, 2012 and June 4, 2013, and classified them into three types based on the following formula:

$$aliveType = \begin{cases} Eraser, & (Now - Old) < 0 \\ Slumber, & (Now - Old) < 2{,}083 \\ Growth, & (Now - Old) \geq 2{,}083 \end{cases}$$

where Now is the number of tweets observed in June 2016 and Old is the number of tweets observed on June 4, 2013. $(Now - Old)$ can obtain the number of posted tweets in the time period between June 5, 2013 and June 19, 2016. Eraser} is a user whose number of tweets in 2016 is fewer than in June 2013. We can guess that they have removed some tweets. Slumber is a user whose growth we could not confirm from June 2013. The threshold value 2,083 for measuring user growth was obtained by the average number of posted tweets of a user per month (57.86) using our dataset. The number of months is 36 from June 2013 to June 2016, and the threshold value is calculated as follows: $36 \times 57.86 = 2{,}083$. Growth is a user whose growth we could confirm by this threshold value. The total number and ratio of users in each alive type are shown in Table 2, based on the above formula. Eraser-type users made up 7.0% and Slumber-type users were 64.5%.

We also showed the number of accounts created each month for alive users in Figure 3. The vertical and horizontal axes are the number of users and the time period from July 2007 to June 2013, respectively. We can observe that the number of created accounts is increasing as time progresses. This agrees with the Twitter Inc. (2014) report published in February 2014. We calculated the threshold value using blue-bar users who created accounts between July 1, 2007 and March 31, 2012 and tallied the alive types using orange-bar users who created accounts between April 1, 2012 and June 4, 2013.

3.2 Feature values

In this section, we explain feature values for analysis. We classified each tweet into three types: tweet, reply, and retweet based on text regular expression. We observed each user's active hours where the numbers of tweets, replies, and retweets were more than one and calculated the average numbers per active hour to measure user behaviors. The list of feature values is shown in Table 3. We can say that these are robust feature values because we can evaluate the amounts of user activity without dependence on beginning time of Twitter use.

Figure 4 shows the numbers of users in each combination of feature values. In all figures, the number of users at each point is shown in a color chart scaled by a common logarithm. The horizontal and vertical axes were rounded down to one decimal place. In Figure 4 (*rp*:*inrp*), we observe the high correlation between *rp* and *inrp* values because many users are mapped at the $rp \approx inrp$ point. On the other hand, in Figure 4 (*rt*:*inrt*), users are widely mapped on a graph. These suggest that since replies have a reciprocal feature similar to conversations, *rp* and *inrp* became nearly equal values for many users, compared with *rt* and *inrt* values.

We can create concrete hypotheses based on these feature values. When *tw* and *rp* are high and low values, he/she doesn't emphasize communication. When *inrp* and *inrt* are high values concurrently, he/she is a high-authority user, e.g. celebrity, because he/she attracts attention from other users.

4. Status and Growth Analysis

4.1 Analyzed Method

In this section, we explain the analyzing method of future user status. Our goal here is to clarify the characteristics to identify growing users. First, we split users into several clusters by feature values using the Gaussian mixture model (GMM), which is an effective soft-clustering algorithm. The

GMM assumes that each data consists of linear overlap of several Gaussian model components. A data vector $\boldsymbol{x}$ is defined by the GMM with K cluster as follows:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k),$$

where π_k is the mixing proportions for cluster k, $\boldsymbol{\mu}_k$ is the mean vector for cluster k, and Σ_k is its standard deviation for cluster k.

Although the GMM needs the number of clusters K as a parameter, we don't know the optimal one in our dataset. Therefore, we employ the Dirichlet process gaussian mixture model (DPGMM) (Gorur 2010), which can automatically decide the optimal number of clusters in addition to GMM function. The cluster number k of each data $\boldsymbol{x}$ is given by the cluster with the maximum mixing proportions π_k.

We can obtain the number of users belonging to each cluster. Our goal here is to clarify the characteristic clusters that easily identify the user status and growth. Therefore, we calculate the point-wise mutual information (PMI) (Church 1990), which measures the ease of occurrence between two independent events. The PMI score $\mathrm{pmi}(k; s)$ between cluster k and status or alive-type s is calculated as follows:

$$\mathrm{pmi}(k; s) = \log_2 \frac{p(k, s)}{p(k)p(s)},$$

where $p(k)$ and $p(s)$ denote the occurrence probabilities of cluster k and status or alive-type s, respectively. $p(k, s)$ is the joint probability of cluster k and status or alive-type s. When $\mathrm{pmi}(k; s)$ is higher than other scores, the cluster \$k\$ and status \$s\$ often co-occur. When $\mathrm{pmi}(k; s)$ is lower than other scores, the cluster k and status s rarely co-occur.

4.2 Results

4.2.1 GMM clustering result

In the result of DPGMM, the optimal number of clusters K was 10 in our dataset. Figure 5 shows the AIC value in each number of clusters in GMM. The horizontal and vertical axes are the number of clusters and the AIC value. Basically, when the AIC value is minimized, the model is accurately built by fewer parameters (Akaike 1998). The vertical and horizontal axes are AIC values and the number of clusters. We can observe that the AIC value decreases as the number of clusters increases. However, the AIC value is sometimes increased since $K = 10$, and the minimum is also achieved at $K = 10$.

The mean feature values and the number and ratio of users in each cluster are shown in Table 4. The status analysis and growth analysis columns shown the users shown in Tables 5 and 6, respectively.

c5 and c7 clusters: The number of users belonging to clusters c5 and c7 in the status analysis dataset is 1. Feature values of $inrp$ and $inrt$ with these clusters are extremely high compared with other clusters. We can guess that both users have high authority in Twitter because they received many replies and retweets per hour. Both clusters are removed in our analysis to evaluate major users in this paper. Both users became Dead status on June 2016. Therefore, neither is contained in the growth analysis dataset.

c1 and c4 clusters: The clusters c1 and c4 are the top and second active users because their numbers of tweets per hour tw is higher than other clusters.

c2 cluster: In both datasets of status and growth analysis, the maximum number of users in all clusters is the cluster c2, all feature values of which are the lowest in all clusters. Especially, they almost never receive replies and retweets from other users $inrp = inrt = 0.0$. We think that they were inactive between April 2012 and June 2013.

c0 and c6 clusters: The communicative feature values of rp, rt, and $inrp$ are similar between clusters c0 and c6. On the other hand, we observe that the cluster c0 received one retweets per active hour $inrt = 1.0$ but the cluster c6 almost never received retweets $inrt = 0.0$.

4.2.2 Status analysis

Table 5 shows the number and ratio of users in each cluster and each status. Each % column shows the ratio of its status in the total number of each cluster. Each PMI column shows the $\mathrm{pmi}(c;s)$ score of its status in this cross-tabulation table.

Dead status: The cluster c0 users seldom became Dead status because their PMI score was the lowest value in all PMI scores. The users with clusters c1 and c4 easily dropped out because their PMI scores were higher than other PMI scores of Dead. As a common feature in both clusters, we can observe that they were the top and second active in all clusters, except clusters c1 and c7.

Lock status: The users belonging to clusters c3 and c9 easily became Lock status. We focus on their $rp \approx inrp$ and $rt \approx inrt$. These results suggest that they emphasize closed bilateral communication with specific users such as friends. The users with clusters c1 and c2 seldom became Lock status. We think that the cluster c1 users deleted their accounts when they stopped using Twitter. On the other hand, the cluster c2 users did not operate their account.

Alive status: We could not find characteristic clusters from PMI scores.

4.2.3 Growth analysis

Table 6 shows the number and ratio of users in each cluster and each alive-type. Each % column shows the ratio of its status in total number of each type. Each PMI column shows the $\mathrm{pmi}(c;s)$ scores of its alive-type in this cross-tabulation table.

Eraser type: The alive users with clusters c1, c4, and c6 often became Eraser type. In our hypothesis for alive users belonging to active clusters c1 and c4, they are Growth type users although they frequently remove tweets because their feature values are higher than other clusters. The alive users with clusters c0 and c2 rarely became Eraser type. As mentioned above, we think that they almost never removed tweets because they did not operate their account.

Slumber type: We could not find characteristic clusters from PMI scores.

Growth type: The alive users with clusters c1 and c6 rarely became Growth type because their PMI scores were lower than other clusters. The characteristic with cluster c6 was $inrt = 0.0$, i.e. they were not retweeted by other users. The clusters c0 and c3 easily became Growth type compared with other clusters. Therefore, we can guess that some users of cluster c0 matured into Growth type for three years. The common features with both clusters were $rp \approx inrp$ and $rt \approx inrt$.

4.3 Findings

Our status analysis found several interesting characteristics for longitudinal status transitions.

1. Active users, e.g. clusters c1 and c4, easily become Dead status after three years and rarely remain Growth type.
2. Users who emphasize bilateral communications by replies and retweets, e.g. clusters c3 and c9, often mature into Growth type although they easily become Lock status after three years.
3. Eraser type users are classified into two major types: active users, e.g. clusters c1 and c4, and users who were not retweeted by other users, e.g. cluster c6.

5. User Growth Prediction

5.1 Prediction methods

In this section, we provide the user growth prediction method based on previous analysis. The prediction strategy is to give a high score to users expected of growth in future. We can preferentially collect tweets in descending order of a user's score. In addition to this, we propose a scoring method for Eraser type users because we deal with the possibility of collecting tweets prior to the deletion of tweets.

Our finding from user status and growth analyses is that active users dropped out more often than inactive users. Even if they were alive status, they often became Eraser type. Furthermore, we found that users who emphasized reciprocal communications using replies and retweets, more easily matured into Growth type than users who engaged in unilateral communications. To detect users who satisfy these two factors, we calculate $Growth$ score as follows:

$$Growth = \frac{1}{tw}\left\{\frac{rp}{(rp - inrp)^2 + 1} + \frac{rt}{(rt - inrt)^2 + 1}\right\},$$

where tw is average number of tweets per hour. rp and $inrp$ denote the number of sent and received replies per hour. When rp and $inrp$ or rt and $inrt$ are nearly equal values, the user engages in reciprocal communications with other users. The users with such a feature are given a high $Growth$ because it lowers $(rp - inrp)^2$ and $(rt - inrt)^2$ in the denominator.

On the other hand, we found that Eraser type users consist of two major groups: active users and users who were not cited in tweets by other users. To achieve these factors, we calculate the $Eraser$ score as follows:

$$Eraser = tw \cdot (rt - inrt)^2.$$

5.2 Experimental evaluations

5.2.1 Evaluation procedure

We prepared the evaluation dataset to measure growth prediction. We attached two labels of “Growth” and “Other” to each user. Users with the “Growth” label are Growth-type in the growth analysis dataset and users with the “Other” label are other users, except for Growth-type users. We calculate the scores explained in the previous section, create user rankings in descending order of each score, and evaluate the effectiveness of each method using precision at K (P@K), which is calculated as follows:

$$P@K = \frac{\#\ of\ growth\ users\ among\ the\ top\ K}{K},$$

in which the top K steps are 10, 30, 50, 100, 200, 500, and 1,000.

As comparison against proposed prediction methods, we use the logistic regression model, which is one of the effective classification models (Cox 1958). We also create rankings based on posterior probabilities obtained by logistic regression. For evaluation of Eraser-type prediction, we prepared labeled datasets and comparison method by the same procedure.

Here, we focus on three components in growth score that consists of $1/tw$, $rp/((rp - inrp)^2 + 1)$, and $rt/((rt - inrt)^2 + 1)$. To examine prediction effectiveness of these three components, we evaluate prediction performance of them by creating rankings of growth users. Similarly, we evaluate prediction performance of two components (i.e., tw and $(rt - inrt)^2$) in eraser score.

5.2.2 Results

Figure 6 shows the precisions of Growth and Eraser in each top K by each scoring method. The vertical and horizontal axes are the precision and the top K, respectively. The weight parameters of each variable estimated by the logistic regression model for Growth and Eraser predictions are shown in Table 7. The highest value in each column is shown in bold.

The most important feature for estimation of Growth and Eraser by logistic regression was the average number of retweets rt per active hour because it is the highest value in all weight parameters.

In Growth estimation precision of Figure 6, the logistic regression model had the maximum precision up to the top 10. Our proposed method showed the maximum precision between the top 30 and 1,000. In Eraser estimation precision, our proposed method achieved the highest score in all of the top K.

Figures 7 and 8 show estimation precision of Growth and Eraser users by each component in Growth and Eraser scores, respectively. In Eraser estimation precision in Figure 8, Eraser score is higher than others between the top 10 and 30. Between the top 30 and 1000, estimation precision with $(rt - inrt)^2$) is similar to Eraser score.

5.3 Discussions

From the results of Figure 6, our proposed method effectively calculated scores for Growth-type users because it was higher precision than the logistic regression model between the top 30 and 1,000. From Table 7, the logistic regression model emphasized the average number of retweets per active hour rt and the average number of tweets per active hour tw. However, our proposed method considered the reciprocity of replies and retweets in each user. We think that one of the users' major motivations for continuing Twitter is communications with other users because they obtain other users' reactions to their own behaviors. Users who engage in reciprocal communication have a growth potential in the future, even if the amount of activity is lower. If the major relationship of a user is unilateral communication with other users, they have the possibility of dropping out even though the amount of activity is greater. On the other hand, our proposed method and logistic regression model together considered the average number of tweets per active hour tw in a negative direction because active users such as clusters c1 and c4 rarely became Growth type from the results of Table 6. We guess that they were classified into Eraser type because they removed more tweets than the number of posted tweets.

From the results of Figure 6, our proposed method appropriately estimated Eraser-type users because it was higher precision than the logistic regression model in all of the top K. From Table 7, the logistic regression model strongly evaluated the average number of retweets per active hour rt and the average number of received replies per active hour $inrp$, compared with other feature values. However, our proposed method used the average number of tweets per active hour tw and the difference value between rt and $inrt$. We think that users who were not retweeted by other users periodically removed their own tweets because they did not obtain any reaction to them by other users.

From the results of Figure 7, our proposed growth score showed higher prediction performance than three components of growth score in all of the top K. Among three components in growth score, $1/tw$ is higher prediction performance compared with others between the top 10 and 100. However, between the top 500 and 1000, both components of $rp/((rp - inrp)^2 + 1)$ and $rt/((rt - inrt)^2 + 1)$ showed higher than $1/tw$. From these results, we think that these components covered different target users. Actually, the correlation coefficients among them (shown in Table 8) was low. Therefore, we believe that the growth score has three components characteristics.

6. Conclusion

In this paper, we analyzed intended user status and growth based on feature values observed in each user behavior. Users were split into several clusters based on five feature values, and we clarified users who easily became each status, using PMI score. To predict growing users over a long time, we proposed a scoring method based on communication reciprocity by replies and retweets.

From our sophisticated experimental evaluations, we found that active users more easily dropped out than inactive users, and users who engaged in reciprocal communications often became Growth type. Also, we found that active users and users who were not retweeted by other users often became Eraser type. Our proposed methods effectively predicted Growth/Eraser-type users compared with the logistic regression model. From these results, we clarified the effectiveness of five feature values per active hour to detect intended Twitter user growth for strategically collecting a large amount of tweets.

In future work, we will demonstrate that Slumber-type users actually grow by aggressive replies and retweets by other users.

References

Akaike, H. (1998), Selected Papers of Hirotugu Akaike, Springer New York, New York, NW.

Aramaki, E.m Maskawa, S., and Morita, M. (2011), "Twitter catches the flu: Detecting influenza epidemics using twitter", In Proceedings of EMNLP2011, Edinburgh, pages 1568-1578.

Bollen, J., Pepe, A., and Mao, H. (2011). "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena", In Priceedings of ICWSM2011, Barcelona, pages 450-453.

Canuto, S., Goncalves, M. A., and Benevenuto, F. (2016), "Exploiting new sentiment-based meta-level features for effective sentiment analysis", In Proceedings of WSDM2016, San Francisco, pages 53-62.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. (2010), "Measuring user influence in twitter: The million follower fallacy", In Proceedings of ICWSM2010, Washington, pages 10-17.

Chalmbers, D., Fleming, S., Wakeman, I., and Watson, D. (2011), "Rhythms in Twitter", In Proceedings of SocialCom2011, Boston, pages 1409-1414.

Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2015), "Antisocial behavior in online discussion communities", In Proceedings of ICWSM2015, Oxford, pages 61-70.

Church, K. W. and Hanks, P. (1990), “Word association norms, mutual information, and lexicography”, *Comput. Linguist.*, Vol. 16 No. 1, pages 22-29.

Cox, D. R. (1958). “The regression analysis of binary sequences (with discussion) “, *Journal Roy Stat Soc B*, Vol. 20 No. 1, pages 215-242.

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C. (2013), “No country for old members: User lifecycle and linguistic change in online communities”, In Proceedings of WWW2013, Rio de Janeiro, pages 307-318.

Dror, G., Pelleg, D., Rokhlenko, O., and Szpektor, I. (2012), “Churn prediction in new users of yahoo! Answers”, In Proceedings of WWW2013, Rio de Janeiro, pages 829-834.

Edwards, J. (2015), “Twitter's user growth problem may be worse than you think”, available at :
http://uk.businessinsider.com/twitter-user-growth-is-worse-than-you-think-2015-2 (accessed 30 April 2017).

Ghosh, R., Surachawala, T., and Lerman, K., (2011), “Entropy-based classification of ’retweeting’ activity on twitter”, paper presented at KDD workshop on Social Network Analysis, 21 August, San Diego, available at: http://arxiv.org/abs/1106.0346 (accessed 30 April 2017).

Gong, W., Lim, E.-P., and Zhu, F. (2015), “Characterizing silent users in social media communities”, In Proceedings of ICWSM2015, Oxford, pages 140-149.

Gorur, D. and Edward~Rasmussen, C. (2010), “Dirichlet process gaussian mixture models: Choice of the base distribution”, *Journal of Computer Science and Technology*, Vol. 25 No. 4, pages 653-664.

Gurajala, S., White, J. S., Hudson, B., Voter, B. R., and Matthews, J. N., (2016), “Profile characteristics of fake Twitter accounts”, *Big Data & society*, Vol. 3 No. 2, pages 1-13.

Java, A., Song, X., Finin, T., and Tseng, B. (2007), “Why we twitter: Understanding microblogging usage and communities”, In Proceedings of WebKDD/SNS-KDD2007, pages 56-65.

Kawale, J., Pal, A., and Srivastava, J. (2009), “Churn prediction in mmorpgs: A social influence based approach”, In Proceedings of CSE2009, San Jose, pages 423-428.

Koh, Y. (2014), “Only 11% of new twitter users in 2012 are still tweeting”, available at :
http://blogs.wsj.com/digits/2014/03/21/new-report-spotlights-twitters-retention-problem/, (accessed 30 April 2017).

Mizunuma, Y., Yamamoto, S., Yamaguchi, Y., Ikeuchi, A., Satoh, T., and Shimada, S. (2014), “Twitter bursts: Analysis of their occurrences and classifications”, In Proceedings of ICDS2014, Barcelona, pages 182-187.

Myers, S. A. and Leskovec, J. (2014), “The bursty dynamics of the twitter information network”, In Proceedings of WWW2014, Seoul, pages 913-924.

Navaroli, N. and Smyth, P. (2015), “Modeling response time in digital human communication”, In Proceedings of ICWSM2015, Oxford, pages 278-287.

Prosser, J. (2016), “Twitter to announce fourth quarter and fiscal year 2015 results”, available at:
https://investor.twitterinc.com/releasedetail.cfm?ReleaseID=948875,
(accessed 30 April 2017).

Rajadesingan, A., Zafarani, R., and Liu, H. (2015), “Sarcasm detection on twitter: A behavioral modeling approach, In Proceedings of WSDM2015, Shanghai, pages 97-106.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010), “Earthquake shakes twitter users: Real-time event detection by social sensors”, In Proceedings of WWW2010, Raleigh, pages 851-860.

Tang, J., Chang, S., Aggarwal, C., and Liu, H. (2015), “Negative link prediction in social media”, In Proceedings of WSDM2015, Shanghai, pages 87-96.

Twitter Inc. (2014), “Twitter reports fourth quarter and fiscal year 2013 results”, available at:
https://investor.twitterinc.com/releasedetail.cfm?ReleaseID=823321
(accessed at 30 April 2017).

Wang, B., Wang, C., Bu, J., Chen, C., Zhang, W.~V., Cai, D., and He, X. (2013), “Whom to mention: Expand the diffusion of tweets by @ recommendation on micro-blogging systems", In Proceedings of WWW2013, Rio de Janeiro, pages 1331-1340.

Yamaguchi, Y., Yamamoto, S., and Satoh, T. (2014), "Behavior analysis methods for twitter users based on transitions in posting activities", International Journal of Web Information Systems, Vol. 10 No. 4, pages 363-377.

Yamaguchi, Y., Yoshida, M., Faloutsos, C., and Kitagawa, H. (2015), "Patterns in interactive tagging networks", In Proceedings of ICWSM 2015, Oxford, pages 513-522.

Yamamoto, S. and Satoh, T. (2015), "Hierarchical estimation framework of multi-label classifying: A case of tweets classifying into real life aspects". In Proceedings of ICWSM2015, Oxford, pages 523-532.

Yamamoto, S., Wakabayashi, K., Kando, N., and Satoh, T. (2016), "Who are Growth Users?: Analyzing and Predicting Intended Twitter User Growth", In Proceedings of iiWAS2016, Singapore, pages 64-71.

Yang, J. and Counts, S. (2010), "Comparing information diffusion structure in weblogs and microblogs", In Proceedings of ICWSM2010, Washington, pages 351-354.

Yuan, N. J., Zhong, Y., Zhang, F., Xie, X., Lin, C.-Y., and Rui, Y. (2016), "Who will reply to/retweet this tweet?: The dynamics of intimacy from online social interactions", In Proceedings of WSDM2016, San Francisco, pages 3-12.

Table 1: The number (#) and ratio (%) of users in each status

Status	#	%
Dead	732,154	21.8%
Lock	459,571	13.7%
Alive	2,160,593	64.5%
Total	3,352,318	100.0%

Table 2: The number (#) and ratio (%) of users in each type

Type	#	%
Eraser	64,543	7.0%
Slumber	602,696	65.4%
Growth	254,195	27.6%
Total	921,425	100.0%

Table 3: Feature values observed in each user

Symbol	Description
tw	Average number of tweets per active hour
rp	Average number of sent replies per active hour
rt	Average number of retweets per active hour
$inrp$	Average number of received replies per active hour
$inrt$	Average number of cited tweets per active hour

Table 4: Mean values and number (#) and ratio (%) of users in each cluster

	Feature values					Status analysis		Growth analysis	
id	*tw*	*rp*	*rt*	*inrp*	*inrt*	#	%	#	%
c0	1.1	0.7	0.7	0.7	1.0	244,047	7.3	51,134	5.5
c1	29.7	21.0	8.6	10.4	10.9	5,988	0.2	2,478	0.3
c2	1.0	0.1	0.4	0.0	0.0	1,319,623	39.4	387,869	42.1
c3	1.9	1.6	1.1	1.6	1.3	398,173	11.9	101,511	11.0
c4	9.5	4.8	2.9	3.5	3.2	78,612	2.3	24,679	2.7
c5	1.0	1.3	1.0	65.4	1950.0	1	0.0	-	-
c6	1.5	0.8	0.5	0.7	0.0	757,678	22.6	234,503	25.5
c7	49.3	66.0	4.9	3914.8	32.3	1	0.0	-	-
c8	3.6	2.6	1.4	2.3	1.7	225,136	6.7	69,258	7.5
c9	1.4	1.3	1.0	1.3	1.1	323,059	9.6	49,996	5.4
Total						3,352,318	100.0	921,425	100.0

Table 5: Number (#) and ratio(%) of users in each cluster and each status

	Dead			Lock			Alive		
id	#	%	PMI	#	%	PMI	#	%	PMI
c0	24,655	10.1	-1.11	31,656	13.0	-0.08	187,736	76.9	0.25
c1	2,276	38.0	0.79	433	7.2	-0.92	3,279	54.8	-0.23
c2	329,952	25.0	0.19	102,760	7.8	-0.81	886,911	68.2	0.06
c3	71,669	18.0	-2.79	89,671	22.5	0.71	236,833	59.5	-0.11
c4	32,400	41.2	0.91	8,517	10.8	-0.34	37,695	48.0	-0.42
c6	167,547	22.1	0.01	107,152	14.1	0.04	482,979	63.7	-0.01
c8	63,931	28.4	0.37	39,677	17.6	0.36	121,528	54.0	-0.25
c9	39,722	12.3	-0.82	79,705	24.7	0.84	203,632	63.0	-0.03
Total	732,152	21.8		459,571	13.7		2,160,593	64.5	

Table 6: Number (#) and ratio (%) of users in each cluster and each alive type

	Eraser			Slumber			Growth		
id	#	%	PMI	#	%	PMI	#	%	PMI
c0	2,185	4.3	-0.71	28,592	55.9	-0.22	20,357	39.8	0.52
c1	281	11.3	0.69	1,821	73.5	0.16	376	15.2	-0.86
c2	18,536	4.8	-0.55	266,218	68.6	0.06	103,115	26.6	-0.05
c3	6,261	6.2	-0.18	52,720	51.9	-0.33	42,530	41.9	0.60
c4	2,517	10.2	0.54	13,637	55.3	-0.24	8,522	34.5	0.32
c6	25,328	10.8	0.62	174,106	74.2	0.18	35,069	15.0	-0.88
c8	6,650	9.6	0.45	35,717	51.6	-0.34	26,891	38.8	0.49
c9	2,776	5.6	-0.33	29,885	59.8	-0.13	17,335	34.7	0.33
Total	64,534	21.8		602,696	13.7		254,195	27.6	

Table 7: Weight values of each feature variable in logistic regression model

	Weight	
Variable	Growth	Eraser
tw	−0. 198	0. 008
rp	−0. 014	0. 003
rt	0. 406	0. 034
$inrp$	0. 139	0. 032
$inrt$	0. 147	−0. 014

Table 8: Correlation coefficients among three components in Growth score

Variable 1	Variable 2	Correlation coefficients
$1/tw$	$rp/((rp-inrp)^2+1)$	−0. 642
$1/tw$	$rt/((rt-inrt)^2+1)$	−0. 392
$rp/((rp-inrp)^2+1)$	$rt/((rt-inrt)^2+1)$	0. 477

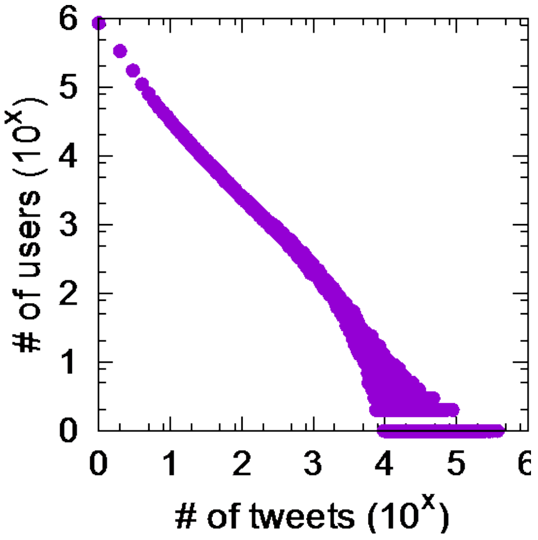


Figure 1: Frequency distribution of users per number of tweets

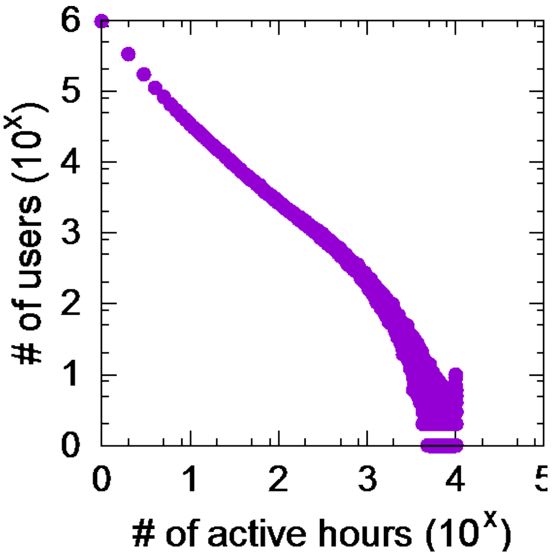


Figure 2: Frequency distribution of users per number of active hours

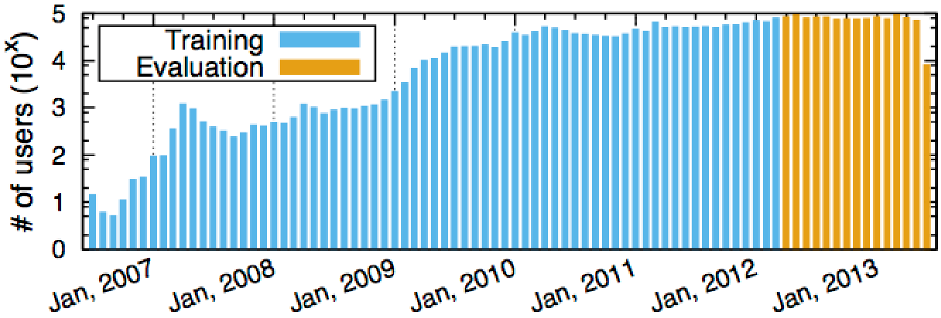
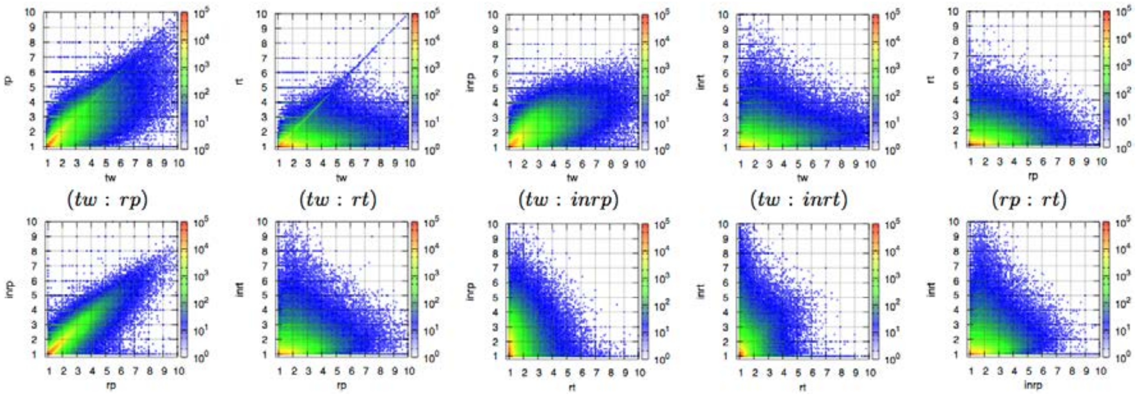


Figure 3: The number of created accounts in each month



$(tw : rp)$ $(tw : rt)$ $(tw : inrp)$ $(tw : inrt)$ $(rp : rt)$

Figure 4: Numbers of users in each combination of feature values

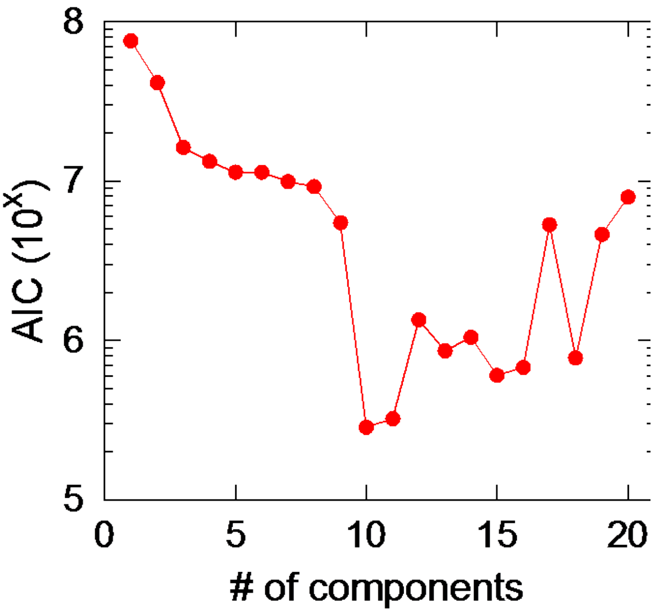


Figure 5: AIC value of each number of clusters in GMM

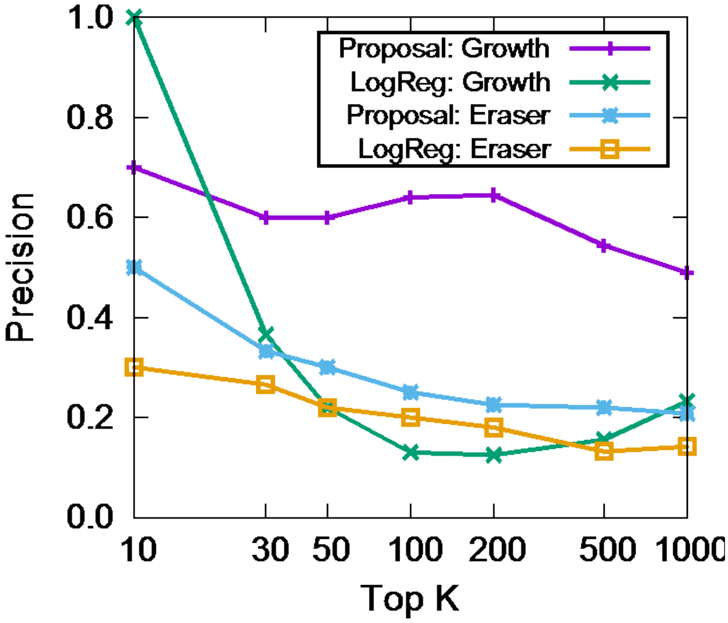


Figure 6: Growth and Eraser prediction precision in each number of top K

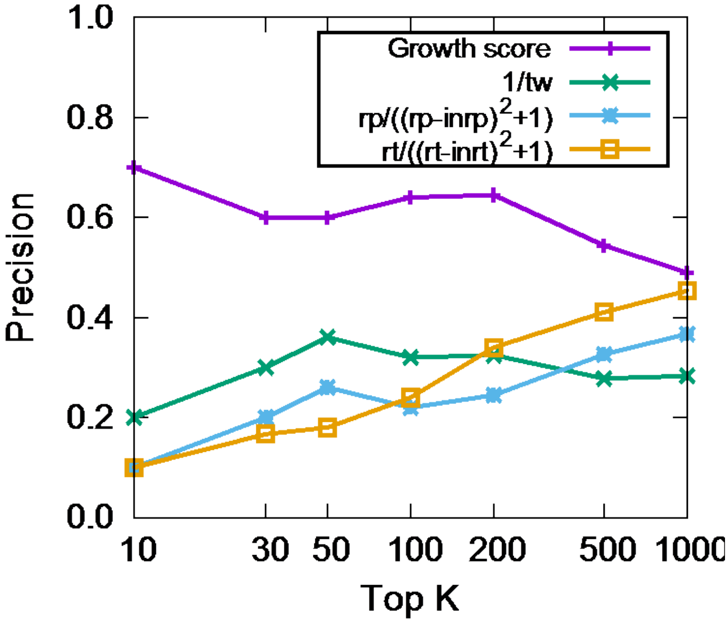


Figure 7: Each component effectiveness of Growth prediction

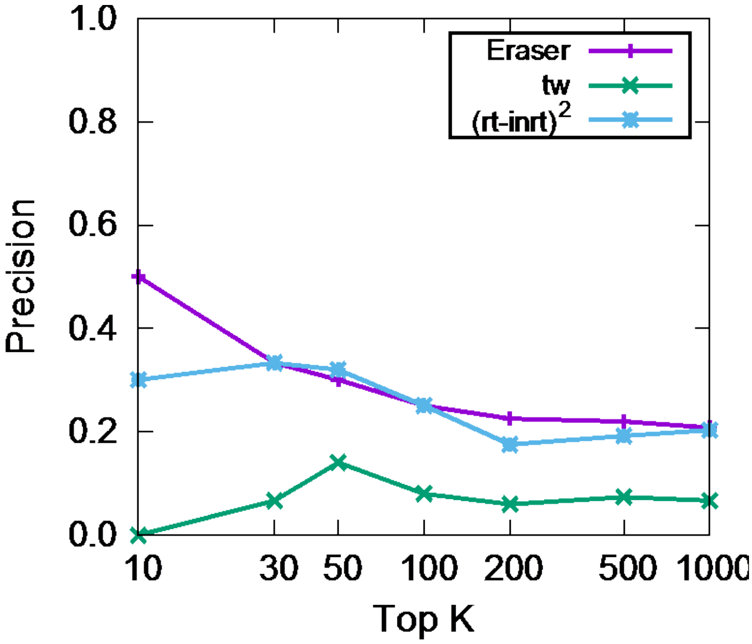


Figure 8: Each component effectiveness of Eraser prediction