Churn Analysis of Online Social Network Users Using Data Mining Techniques

Xi Long†, Wenjing Yin, Le An, Haiying Ni, Lixian Huang, Qi Luo, and Yan Chen

***Abstract*—A churn is defined as the loss of a user in an online social network (OSN). Detecting and analyzing user churn at an early stage helps to provide timely delivery of retention solutions (e.g., interventions, customized services, and better user interfaces) that are useful for preventing users from churning. In this paper we develop a prediction model based on a clustering scheme to analyze the potential churn of users. In the experiment, we test our approach on a real-name OSN which contains data from 77,448 users. A set of 24 attributes is extracted from the data. A decision tree classifier is used to predict churn and non-churn users of the future month. In addition, *k*-means algorithm is employed to cluster the actual churn users into different groups with different online social networking behaviors. Results show that the churn and non-churn prediction accuracies of $\sim$65% and $\sim$77% are achieved respectively. Furthermore, the actual churn users are grouped into five clusters with distinguished OSN activities and some suggestions of retaining these users are provided.**

***Index Terms*—Online social network, churn prediction, user clustering, retention solution.**

I. Introduction

WED-BASED online social network (OSN) sites have large beneficial effects on people's social communications through internet [1], [2]. Many OSNs (e.g., Facebook, LinkedIn, and MySpace), which provide various services to meet different user needs, are emerging and growing rapidly with easier and better interactive experience [3]. A large number of netizens register their personal accounts at OSN sites with social communities in order to extend their social relationships in local, national or worldwide ranges by connecting with others. Such connections between individuals are usually based on their manifest or latent relationships fashioned by their self-disclosed information such as personal profiles, locations, and interests [4]. Up to now, more than a hundred OSNs have been launched and many of them focus on providing local services to users in their own countries with mother-tongue-only interfaces by means of their inherent advantages in the aspects of local culture, daily lifestyle, online/offline behavior, communication habit, and so on [5]. For example, in China, the most widely-used Chinese OSNs are: Renren (renren.com), Qzone (qzone.qq.com), Weibo (weibo.com), Kaixin001 (kaixin001.com), and Pengyou (pengyou.com), who share the majority of China's OSN market, traffic, and users [5], [6].

Many studies based on OSN users with different purposes have been achieved by using data mining techniques [7]-[9]. For an OSN service provider, the sustainable development of the OSN site mainly depends on whether it has mass users and how active they are [4]. Unfortunately, a problem has been observed that many existing users behave inactively on an OSN site or they completely leave the site. In other words, the OSN site loses users. In this paper, the loss of a user is defined as a churn. The analysis of user churn problem has been studied in many different industries such as telecommunications [10], [11], healthcare [12], financial service [13], and banking [14]. For OSN service providers, it has become increasingly important because user churn of OSN sites causes direct revenue loss [15]. Besides, recruiting a new user may cost higher than retaining an existing one in internet industry [2], [15]. Therefore, it is important for an OSN service provider to: (1) precisely and early locate the users who are at high risk of churn in the near future (i.e., the future month in this case); (2) timely prevent these users from churning by using appropriate retention solutions such as providing interventions through email, offering better services, recommending interesting contents, etc.

Although the conventional approaches (e.g., usability test and user interview) are useful for qualitatively knowing the interactive problems and user expectations of an OSN site, they are less helpful to identify who may churn in the future. Hence, predicting future churns based on their past online activity data plays an important role in delivering appropriate retention solutions to them. In this study, the prediction was achieved by means of a classifier that can distinguish the users in two pre-defined classes (i.e., churn and non-churn) because their online activity data contain information from which churn and non-churn users can be derived. Classifiers require that this information is first extracted from the data as "attributes". Several simple classifiers can be considered for classifying churns and non-churns with acceptable performances such as neural network (NN) and decision tree (DT) [16]. Compared with a DT-based classifier, a NN-based one cannot give an explicit expression of the attribute patterns of classes in an easily understandable form, which complicates user behavior interpretation for making decisions during classification [17]. It means that only the churning probabilities of users are provided without indicating which attributes are more significant in discriminating with churn and non-churn. Besides, a DT-based

Manuscript received December 08, 2011. This work was supported by Customer Research and User Experience Design Center (CDC), Tencent Inc., China.

†X. Long was a User Researcher in CDC, Tencent Inc., High-Tech Park, Shenzhen, 518057, China. He is now a PhD candidate in the Dept. of Electrical Engineering, Eindhoven University of Technology, Eindhoven, 5600 MB, the Netherlands (e-mail: xi.long.ee@gmail.com).

W. Yin[1], H. Ni[2], L. Huang[3], and Q. Luo[4] are Senior User Researchers in CDC, Tencent Inc., High-Tech Park, Shenzhen, 518057, China (e-mail: {[1]viviyin, [2]amieeni, [3]henryhuang, [4]sybil}@tencent.com).

L. An is a PhD candidate in the Dept. of Electrical Engineering, University of California, Riverside, CA92521, USA (e-mail: lan004@ucr.edu).

Y. Chen is a Director in CDC, Tencent Inc., High-Tech Park, Shenzhen, 518057, China (e-mail: enya@tencent.com).

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

classifier [18] is adopted to predict the class membership.

After locating the churn users, it is important to group these users into different clusters based on their different online activities so that the appropriate retention solutions can be delivered to them. For this purpose, unsupervised learning is recommended. From the practical point of view, a *k*-means and a hierarchical clustering algorithm are considered due to the wide employment and the ease of implementation. However, a hierarchical algorithm requires a quadratic computation complexity of O(N^2) which is very intensive to be applied on an internet dataset with a large number of observations, while *k*-means algorithm only requires a computation complexity of O(N) [19]. So a *k*-means clustering algorithm is employed in this study for grouping the churn users. Furthermore, after clustering process, a "second-step" classifier can be determined based on the clustering results that serve as the "ground-truth". This classifier aims at classifying an incoming user, who has been predicted as a churn, into one of the defined clusters. Then the most appropriate retention solution can be precisely delivered to this user. This way is more effective to motivate churn users to keep on using an OSN site because providing all registered users a similar solution fails to consider that they usually have different needs, intentions, and interests on the site.

The main contributions of this paper include: first, a large real-world dataset is collected to study the churn pattern in real application; second, a framework of churn prediction is developed; third, we cluster the churn users based on their online activities; fourth, different solutions of preventing users from churning are suggested to different clusters.

II. Data

Pengyou (formerly Xiaoyou), literally "friend", is one of the famous OSN sites of China. It has been reported that Pengyou has more than a hundred million registered users [20]. Pengyou is a derivative product of an instant messaging (IM) client of Tencent Inc., generally referred as QQ (imqq.com), which is the most widely-used and popular IM software in China with more than a billion registered users [21]. In fact, a "QQ-connection strategy" allows Pengyou users logging in the site using their emails or QQ accounts. It means that it allows a user automatically adding all QQ contacts in Pengyou as friends. Thus, Pengyou shares a large portion of the population of QQ users, which indicates that a user is able to access both services with a unique registered account. In addition, Pengyou has a function of so-called "QQ synchronization". For instance, some new feeds in Pengyou (e.g., messages, comments, statuses, and friend invitations) are allowed to, after being chosen yes to synchronize, be exported to a user via his/her QQ in the form of reminder without any delay and obstacle, and vice versa.

In our experiment, the online activity data within a period of one month (i.e., October, 2008) were collected from the users on Pengyou. They were annotated to be churn or non-churn by examining login activities to the site in the future month (i.e., November, 2008). In other words, churn is the case that there is no login activity to Pengyou in the future month whereas there still is login activity in the current month. The users in such case are labeled as "churn" while the others are labeled as "non-churn". In this study, the online activity data from 100,000 users (i.e., instances) were collected in anonymity (in honor of user privacy and under certain regulations) such as number of friends, total online time, login times, active days, number of blogs written, number of pictures shared, number of messages in guestbook, login times in QQ, etc. Then a large dataset was built in which various attributes (e.g., users' statuses and login activities) can be extracted. Note that all the data are monthly-based. For instance, the number of monthly login times is computed by summing the number of daily login times of all the days in this month. Details of the attributes will be given in Section III. Here the demographic information of users like gender and age are excluded because such information may be untrue or incompletely provided by users due to their consideration of security and privacy.

In addition, after reviewing the dataset, some values of the data perform very high or low comparing with others. These outliers might be due to the system mistake when data were collected. In particular, a small portion of the values of some attributes (e.g., the number of playing online games on Pengyou) are extremely high. It is because some online game addicts used hack tools for cheating on games on this OSN site. Since including these erroneous data seriously contaminates the population and may give invalid results, the original dataset should be cleaned up. The hacker-generated outliers can be identified and then be discarded based on the rules of hacker data detection. For the rest un-hacked data, the threshold of detecting outliers is different for each attribute. A conventional three-sigma test [22] can remove the outliers lying outside the three times standard deviations of the mean of an attribute's values. This empirical test assumes that the data under test should be normally distributed, which is suggested by using a *Q-Q* plot method. Thus, a clean dataset with 77,448 instances without outliers is used for training and testing of a classifier. A portion of the clean dataset with 25,170 churn instances ($\sim$32%) is used for building a clustering model.

III. Attribute Extraction

In total, 24 attributes are extracted from the data within a month and are divided into five groups. They are in the aspects of *1) status*, *2) login activity*, *3) basic operation activity*, *4) App (application) operation activity*, and *5) IM activity*. Note that the values of each attribute are normalized. A brief description of the five groups follows and the attributes are listed in Table I.

1) Status Attribute - the number of friends of a user on the OSN site indicates to what extent a user is connecting with others.

2) Login Attribute - the monthly web-access activity in Pengyou reflects that how active a user is in the community. It can be measured by the attributes of monthly numbers of login times, login days, active times, and active days. Number of login times counts the number of logins to the OSN site and number of login days sums the days in which login occurs. A specific threshold is used to label if a user performs active or inactive during the time interval from each login to then log-off. Number of active times sums the number of active time intervals of a month and number of active days records the number of days, in one of which a user performs active at least once. Total online time means the total time (in minutes) of a user being online in a month.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE I
LIST OF ATTRIBUTES

Group	Attribute
Status attribute	number of friends
Login attributes	number of login times number of login days number of active times number of active days number of total online time
Basic operation attribute	number of blogs written number of videos shared number of pictures uploaded number of friend invitations sent/received number of messages in guestbook number of times of status updated number of times of events updated number of times of privacy settings edited number of times of profile updated total number of basic operations
App operation attribute	number of times of playing game 1 number of times of playing game 2 number of times of playing game 3 total number of App operations
IM attributes	number of login times number of login days number of messages membership level

3) Basic Operation Attribute - exploring the operations of playing basic functions of the OSN site provides more details in understanding to what extent a user is attracted by these functions. Attributes can be extracted based on the operations such as number of blogs written, number of videos shared, number of pictures uploaded, number of friend invitations sent and received, number of messages in guestbook, number of times of status updated, number of times of events updated, number of privacy settings edited, number of times of personal profile updated, and the total number of these basic operations.

4) App Operation Attribute - rather than basic operation attributes, applications seem more attractive to many users due to their fun of playing. Currently, releasing application programming interfaces (APIs) to the third parties provides opportunities to a social network platform on building a more active and enjoyable environment on the site. Therefore, such applications often imply how active an OSN is. Here the number of times of playing an online game application is considered and three featured games are involved, which means three attributes are extracted. Besides, the total number of times of playing these three games is also included.

5) IM Attribute - as explained in Section II, a user is allowed to log in the OSN site (i.e., Pengyou) and the IM client (i.e., QQ) by using the same username and password. So such connection and synchronization of these two platforms indicate that the activity of using the OSN site may highly correlate to that of using the IM client, which thus is worthwhile to be investigated. Therefore, the number of login times and the number of login days on QQ are extracted. Number of messages includes the total number of messages sent and received in that month. And membership level is proportional to the total time and activity of using QQ.

IV. TECHNICAL APPROACH

In order to provide timely and effective retention solutions to prevent users from churning in the future month, the users should be first detected before they actually churn.

A. Prediction

A DT-based classifier is adopted in this study to predict churn and non-churn users. It is a simple yet powerful method for deriving classification rules from a set of labeled instances [18]. Over many other classifiers, a DT-based classifier has advantages that, first, it can be extended to non-numeric domains where the attributes are categorical data rather than numerical data; second, a decision made from the node to the final labeling is easy to follow and interpret. It helps to understand attributes that perform as better indicators in predicting churn instances. This delivers a useful message of making a proper retention solution to prevent users from churning. However, tuning parameters of the tree such as the numbers of leaves and nodes should be careful in order to avoid over-fitting (or under-fitting) the data. A DT-based classifier is achieved by making a sequence of decisions along a path of the nodes of a constructed tree trained by the attributes of training set. In the training phase, at each node of the tree, a decision rule splits the instances into two or more partitions and new nodes are then generated. Splitting criterion is adopted according to which the best split from the set of candidates is chosen. The stopping criterion controls the growth of the tree and a node is considered to be terminal or a leaf node. Once a tree is built after training, new instances are able to be classified via making a sequence of decisions on their corresponding attributes. More details of the DT-based method can be found in [18] and [23]. Usually, overall accuracy, which is computed through dividing the number of instances by those correctly predicted, is used to measure the performance of classification. This agrees with the assumption that, during the decision making step, the same misclassification costs are assigned to all classes. However, such overall accuracy may not be a perfect criterion for classification, especially for the situation of unbalanced class distribution. For the binary classification problem in this study, a well refinement was achieved by means of receiver operating characteristic (ROC) analysis, which can adjust the global decision making threshold to be optimal [24]. It is biased to search the optimal threshold by maximizing the sum of sensitivity and specificity in plotted ROC curve, which is one of the most commonly used criteria, known as Youden's criterion [25]. For any given threshold T, the Youden's index is

$$Y = \arg\max_{T}(SE(T) + SP(T) - 1) \qquad (1)$$

where SE is sensitivity (i.e., true positive rate) and SP is specificity (i.e., true negative rate) in a 2×2 confusion matrix in a binary-class problem (see Table II).

B. Clustering

After the users with high risk of churn are detected, it is important to analyze these users in order to further provide them with the most appropriate solutions of retention. Here unsupervised learning is employed to group the churn instances. The predicted churn users who are not actual churns

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE II
CONFUSION MATRIX OF A TWO-CLASS PROBLEM

	Predicted churn	Predicted non-churn
Actually churn	True Positive	False Negative
Actually non-churn	False Positive	True Negative

are rejected in the learning process because including these misclassified instances may lead to bad performance of a clustering model.

In this study, a *k*-means clustering algorithm [26] is used. The main purpose of *k*-means algorithm is to define k centroids, one for each cluster. The number of clusters k is known *a priori* and the centroids are estimated by means of iterative refinement process. In the beginning, k points are randomly chosen as the initial points to create k clusters, which are partitioned by associating those points with nearest means in the corresponding clusters. Then the cluster means become the new centroids. Such processes are repeated until dynamic centroids reach convergence, which means the locations of the k centroids have no more changes. The final centroids of the k clusters are then determined. Here given a set of observations $x_1, x_2, \ldots, x_n$, where each of them is a user's attribute vector with m dimensions. Here n is the number of instances in the dataset and m is the number of extracted attributes. The *k*-means algorithm partitions the n instances into k groups $G = G_1, G_2, \ldots, G_k$ by minimizing the following function within clusters:

$$J = \arg\min \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - c_j\|^2. \quad (2)$$

where c_j is the mean of points in cluster j denoted as the cluster center, $\|x_i^{(j)} - c_j\|^2$ is the square of Euclidean distance between a data point and the cluster center c_j.

Once the centroids of clusters are defined, a new incoming user is classified by selecting the minimal Euclidean distance between this sample and the previously defined centroids. The question then is that how to select the number of clusters k $(0 \leq k \leq n)$. Here we use a selection method based on clustering results [27], where an objective evaluation measure is employed to select the number of k. The measure is calculated as a distortion function $DF(k)$ that converges to a constant value when k increases. The minimum of $DF(k)$ is suggested where k should be regarded as the number of clusters. The distortion function $DF(k)$ is given by:

$$DF(k) = \begin{cases} 1, & if\ k = 1 \\ \frac{D_k}{\alpha_k D_{k-1}}, & if\ D_{k-1} \neq 0, \forall k > 1 \\ 1, & if\ D_{k-1} = 0, \forall k > 1 \end{cases} \quad (3)$$

where k is the number of clusters, D_k is called the sum of all distortions. The distortion of cluster j is the square sum of all the distances between within-cluster data points and the cluster center in this cluster. D_k is computed as follow:

$$D_k = \sum_{j=1}^{k} \sum_{l=1}^{n_j} \langle x_{j,l}, c_j \rangle^2 \quad (4)$$

where $\langle x_{j,l}, c_j \rangle^2$ is the distance between the lth data point $x_{j,l}$ and the center c_j of cluster j, and n_j means the number of data points within cluster j. Besides, α_k is a weight factor that equals to $1 - 0.75d$ when $k = 2$ and $\alpha_{k-1} + (1 - \alpha_{k-1})/6$ when $k > 2$, here d is the number of attribute dimensions in the dataset. The evaluation function aims at selecting optimal k by minimizing the sum of within-cluster distortions.

V. RESULTS AND DISCUSSION

As observed from the annotation obtained in the future month, the number of churn users takes ∼32% in the clean dataset. A 10-fold cross validation was used in the experiments. Fig. 1 shows the ROC space of the prediction performance by using a DT classifier, where the optimal searched result is marked based on the sweep of a decision making threshold in terms of Youden's criterion. By applying ROC optimization in the classification, the Youden's index reached 0.42 with an obvious improvement of ∼0.06 compared to the original result. Table III shows an increase of 16.6% in sensitivity (i.e., churn prediction accuracy) and a comparably less decrease of 10.8% in specificity (i.e., non-churn prediction accuracy), even there is a slightly decrease in overall accuracy. Among all the extracted attributes, the one number of friends achieved the strongest capability in discriminating with churn and non-churn users on Pengyou. Some other classifiers (e.g., *k*-nearest-neighbors, naïve Bayes) have also been examined, but they offered no significant improvement compared with a DT-based classifier.

As discussed in Section II, 25,170 users' data in the clean dataset were taken into account for building a clustering model. Fig. 2 shows that the number of the clusters k is preselected to be 5 as the best one based on the clustering performances by sweeping the k from 1 to 20. In practice, selecting a larger k makes no sense that is considered to be overfitting the clustering model. It also complicates interpreting the reasons of user churn and delivering retention solutions to more than 20 clusters. Fig. 3 indicates the distribution of the five distinguished churn clusters. In terms of the OSN activities of the churn users on Pengyou in the past month, some details of the five clusters are given in the Table IV, where five most representative attributes are visualized for comparison between the average of all the churn users and the average of the churn users belonging to a specific cluster. Hence, relative difference ratio (RDR) is measured for such comparison. Followings are the explanations the five churn clusters with different online activities.

• Cluster 1 ("*come-and-leave users*") - Approximately 30% of the churn users are observed in this cluster who generally behaved inactively. They did not have many friends in the community. They were with less logins both on Pengyou and QQ and were below average in both basic and App operations in the past month. A reason might be that such users registered Pengyou accounts by accident but they could not find any interesting things before they left. Thus this cluster of users can be named by "*come-and-leave users*".

• Cluster 2 ("*IM-active-only users*") - The users in this cluster were inactive on Pengyou, rather that they were quite active on QQ. It can be imagined that these users might come to Pengyou via "QQ synchronization" but they still preferred to stay in their QQ networks. Another reason might be that Pengyou, as an social network site, could not be clearly differentiated from a traditional instant messaging network.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

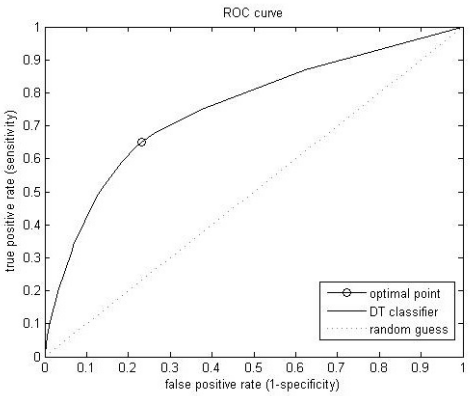

Fig. 1. Prediction performance using a DT classifier in the ROC space. The optimal result, represented by the circle marker, is searched based on Youden's criterion. The dotted line represents the classes with no discrimination as a random guess.

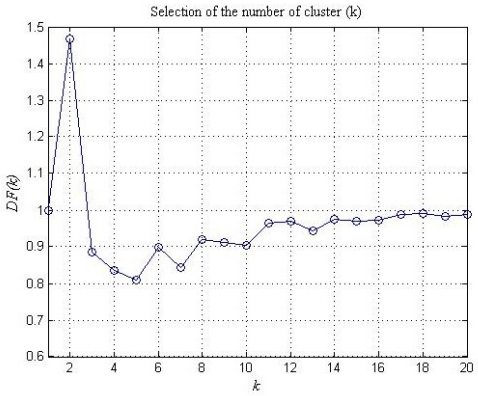

Fig. 2. Evaluation of $DF(k)$ by sweeping k from 1 to 20 for k selection.

TABLE III
COMPARISON OF PREDICTION PERFORMANCE

	Original Results (without ROC)	Optimal Result (with ROC)
Sensitivity	48.5%	65.1%
Specificity	87.6%	76.8%
Overall accuracy	74.9%	73.0%
Youden'index	0.36	0.42

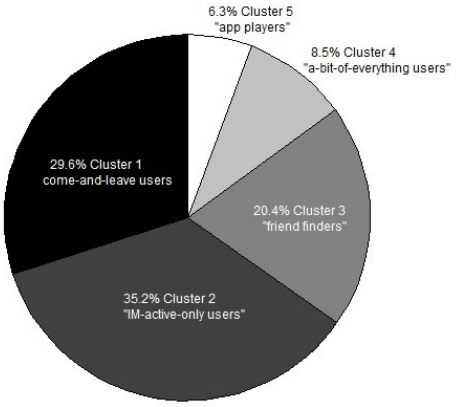

Fig. 3. Distribution of the five churn clusters.

• Cluster 3 ("*friend finders*") - Friend finders usually focus on seeking their old friends or making new friends through an open social network. They can realize that an OSN site offers them a more convenient approach where they are able to connect with others for socializing on internet. However, they churned eventually even they have many friends at hand and are interested in interacting with their friends. It might be due to that their friends were not the ones as expected or these friends were not active on the site. Another explanation is that these users only collected friends rather to communicate with them. "*friend finders*" takes ∼20% of the total churn users.

• Cluster 4 ("*a-bit-of-everything users*") - It is noticed that a cluster of churn users (∼9%) behaved active in basic operations as well as in logging in Pengyou. They did not often play OSN Apps or games, but preferred to use around the typical OSN functions such as sharing contents, writing blogs, updating status, uploading pictures, and etc. However, this cluster of users left the site after all, which might result from that the basic functions were not sufficient to make them stay in the network for a longer time. In other words, they could only realize this is a "website" rather than a "network".

• Cluster 5 ("*App players*") - There are around 6% of the churn users who were only fascinated in playing OSN-based game applications. They were completely attracted by games rather than any other functions or social activities. Hence, these users might leave the site in case of feeling jaded or bored of a game because of the "life-cycle limit" of an online game/App for players.

The OSN activities are qualitatively summarized in Table V with the denoted levels of "high", "moderate", and "low". Actually, in the clustering process, the attributes can be weighted. The weights usually depend on the preferences and/or the specific requirements of investigation of an OSN service provider. They may also depend on that in which aspects, corresponding to attributes, are being focused more. But these weights need to be carefully determined in practice. In terms of the characteristics of the five clusters, specified retention solutions should be carefully studied for the purpose of effectively preventing users from churning. For example, for the "*come-and-leave users*" and "*a-bit-of-everything users*", user interview and focus group might be good approaches to use in order to understand user needs and interests on an OSN site. It is because a qualitative investigation usually yields good explanations to the results of a quantitative study. Recommending new friends who are active on the site to the "*friend finders*" is useful to prevent them from churning. It is also important to design a better user interface for them to find these new friends more easily. It has been observed that "*App players*" prefer online games to others on an OSN site. Then, developing essential Apps for them fits their needs well. A reason of their churning might be due to that the life-cycles of the games they played are too short. So automatically collecting many new Apps for them according to their preferences can be an effective solution of retention. The problems such as "what kind of Apps a user like" and "in what context he/she plays" are suggested to be further investigated. For the "*IM-active-only-users*", apparently, it may be helpful to intervene through the synchronized instant messenger (i.e., QQ in this case), such as providing a quick entrance or interesting feeds of Pengyou in the client panel of the instant messenger.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE IV
THE RELATIVE DIFFERENCE RATIOS (RDRS) OF THE FIVE REPRESENTATIVE ATTRIBUTES OF THE CHURN USERS IN THE FIVE CLUSTERS

	Number of friends	Number of OSN login times	Total number of basic operations	Total number of App operations	Number of IM login times
Cluster 1 ("*come-and-leave users*")	-0.5	-0.4	-0.3	-0.5	-0.6
Cluster 2 ("*IM-active-only users*")	-0.5	-0.2	-0.3	-0.7	0.6
Cluster 3 ("*friend finders*")	1.5	-0.5	0.0	-0.8	-0.2
Cluster 4 ("*a-bit-of-everything users*")	0.1	0.6	2.0	-0.4	-0.1
Cluster 5 ("*App players*")	-0.2	4.6	0.1	10.2	0.2

TABLE V
QUALITATIVE SUMMARY OF CHURN USERS' OSN ACTIVITIES WITH DENOTED LEVELS "HIGH", "MODERATE", AND "LOW"

	Status (nr. of friends)	Login activity	Basic operation activity	App operation activity	IM activity
Cluster 1 ("*come-and-leave users*")	Low	Low	Moderate	Low	Low
Cluster 2 ("*IM-active-only users*")	Low	Moderate	Moderate	Low	High
Cluster 3 ("*friend finders*")	High	Low	Moderate	Low	Moderate
Cluster 4 ("*a-bit-of-everything users*")	Moderate	High	High	Low	Moderate
Cluster 5 ("*App players*")	Moderate	High	Moderate	High	Moderate

VI. CONCLUSION

In this paper, we propose an approach to reveal the churn pattern of the OSN users using data mining techniques. We collected a large clean dataset of 77,448 users from one of the largest OSN sites in China, called Pengyou. A DT-based and a *k*-means algorithms allow us to process large amount of data in a timely manner. The experiment shows that $\sim$77% non-churn users and $\sim$65% churn users are correctly predicted in terms of Youden's criterion rather than overall accuracy. In addition, a clustering model is built based on the 25,170 actual churn users, which are grouped into five clusters by using a *k*-means clustering algorithm. The churn users of each cluster are analyzed and some suggestions of retention solutions are provided to prevent them from churning. In other words, by analyzing the known online activities of users in the current month, they are first detected to be churn or non-churn users of the future month and then are classified into one of the five clusters if they are likely to churn. Consequently the most appropriate retention actions can reach them in an early stage.

REFERENCES

[1] D. M. Boyd and N. B. Ellison, "Social network sites: definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 1-19, 2007.
[2] E. Hargittai, "Whose space? Differences among users and non-users of social network sites," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 276-297, 2007.
[3] W. Duan, "Special issue on online communities and social network: an editorial introduction," *Decision Support Systems*, vol. 47, pp. 167-168, 2007.
[4] Y. Jia, Y. Zhao, and Y. Lin, "Effects of system characteristics on users' self-disclosure in social networking sites," in: *7th Int. Conf. Information Technology*, 2010, pp. 529-533.
[5] Alexa Web Information Company. Available: http://www.alexa.com.
[6] K. Lukoff, "China's top 15 social networks," *TechRice*, 2011. Available: http://techrice.com/2011/03/08/chinas-top-15-social-networks.
[7] L. Guo, *et al.*, "Analyzing patterns of user Content generation in online social networks," in: *Proc. 15th Int. Conf. ACM SIGKDD*, 2009, pp. 369-378.
[8] D. Jensen and J. Neville, *Data mining in social networks - Dynamic Social Network Modelling and Analysis: Workshop Summary and Papers*, National Academy of Science Press, pp. 289-302, 2003.
[9] R. Baden, *et al.*, "Persona: an online social network with user-defined privacy," in: *Proc. ACM SIGCOMM*, 2009.
[10] J. Lu, "Predicting customer churn in the telecommunications industry - an application of survival analysis modeling using SAS?," in: *Online Int. Proc. of SAS User Group*, 2002, paper 114-27.
[11] S. Y. Hung, D. C. Yen, and H. Y. Wang, "Applying data mining to telecom churn management," *Expert Systems with Applications*, vol. 31, no. 3, pp. 515-524, 2006.
[12] M. Pijl, *et al.*, "Prediction of successful participation in a lifestyle activity program using data mining techniques," in: *21st Benelux Conf. Artificial Intelligence,*, 2009.
[13] J. Burez and D. Van den Poel, "Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department," *Expert Systems with Applications*, vol. 35, no. 1-2, pp. 497-514, 2008.
[14] Y. Xie, *et al.*, "Customer churn prediction using improved balanced random forests," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5445-5449, 2009.
[15] M. Karnstedt, *et al.*, "Part I: Social media analysis and organization - churn in social networks," it Handbook of Social Network Technologies and Applications, Springer, pp. 185-222, 2010.
[16] J. Ferreira, *et al.*, "Data mining techniques on the evaluation of wireless churn," in: *European Sym. Artificial Neural Networks*, pp. 483-488, 2004.
[17] W. H. Au, K. C. C. Chan, and X. Yao, "A novel evolutionary data mining algorithm with applications to churn prediction," *IEEE Trans. Evolutionary Computation*, vol. 7, no. 6, pp. 532-545, 2003.
[18] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81-106, 1986.
[19] E. Keogh and J. Lin, "Clustering of time-series subsequences is meaningless: implications for previous and future research," *Knowledge and Information Systems*, vol. 8, no. 2, pp. 154-177, 2005.
[20] S. Ye, "Pengyou: Tencent's latest real name social network," *TechRice*, 2011. Available: http://techrice.com/2011/01/05/Pengyou-tencents-latest-real-name-social-network.
[21] Tencent Inc. official website. Available: http://www.tencent.com/en-us/at/roadmap.shtml.
[22] D. Ruan, G. Chen, E. E. Kerre, and G. Wets (eds.), "Intelligent data mining: techniques and applications," *Studies in Computational Intelligence*, Springer, vol. 5, pp. 318, 2005.
[23] S. Theodoridis and K. Koutroumbas. *Pattern Recognition* (2nd ed.), Elsevier Academic Press, 2003.
[24] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowledge and Data Engineering*, vol. 17, pp. 299-310, 2005.
[25] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, pp. 32-35, 1950.
[26] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in: *Proc. 5th Berkeley Sym. Mathematical Statistics and Probability*, University of California Press, vol. 1, pp. 281-297, 1967.
[27] D. T. Pham, S. S. Dimov, and C. D. Nguyen, "Selection of *k* in *k*-means clustering," *Mechanical Engineering Science*, vol. 219, pp. 103-119, 2004.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)