Understand and Assess People's Procrastination by Mining Computer Usage Log

Ming He[1], Yan Chen[1], Qi Liu[1], Yong Ge[2], Enhong Chen[1](✉), Guiquan Liu[1], Lichao Liu[3], and Xin Li[4]

[1] University of Science and Technology of China, Hefei, Anhui 230026, China
heming01@foxmail.com, ycwustc@mail.ustc.edu.cn,
{qiliuql,cheneh,gqliu}@ustc.edu.cn

[2] Nanjing University of Finance and Economic, Nanjing, China
ygestrive@gmail.com

[3] IBM (China) Investment Co Limited, Beijing 100193, China
llcliu@cn.ibm.com

[4] iFlyTek Research, Hefei, China
xinli2@iflytek.com

Abstract. Although the computer and Internet largely improve the convenience of life, they also result in various problems to our work, such as procrastination. Especially, today's easy access to Internet makes procrastination more pervasive for many people. However, how to accurately assess user procrastination is a challenging problem. Traditional approaches are mainly based on questionnaires, where a list of questions are often created by experts and presented to users to answer. But these approaches are often inaccurate, costly and time-consuming, and thus can not work well for a large number of ordinary people. In this paper, to the best of our knowledge, we are the first to propose to understand and assess people's procrastination by mining user's behavioral log on computer. Specifically, as the user's behavior log is time-series, we first propose a simple procrastination identification model based on the Markov Chain to assess user procrastination. While the simple model can not directly depict reasons of user procrastination, we extract some features from computer logs, which successfully bridge the gap between user behaviors on computer and psychological theories. Based on the extracted features, we design a more sophisticated model, which can accurately identify user procrastination and reveal factors that may cause user's procrastination. The revealed factors could be used to further develop programs to mitigate user's procrastination. To validate the effectiveness of our model, we conduct experiments on a real-world dataset and procrastination questionnaires with 115 volunteers. The results are consistent with psychological findings and validate the effectiveness of the proposed model. We believe this work could provide valuable insights for researchers to further exploring procrastination.

W. Liu et al. (Eds.): KSEM 2018, LNAI 11061, pp. 187–199, 2018.
https://doi.org/10.1007/978-3-319-99365-2_17

1 Introduction

With the rapid growth of Information Technology, computer and Internet have been playing an important role in people's daily life. We use softwares on computer and internet to conduct daily work, and meanwhile we also play games or browser online content a lot. Although the computer and Internet bring our lives great advantages, they also result in various problems to our work. Particularly, the unlimited access to Internet bring us much distraction, e.g., procrastination. Procrastination, which is a general issue of people in modern life, is to voluntarily delay an intended course of action despite expecting to be worse off for the delay [17].

As the expansion of emerging procrastination, it has attracted much researchers' attention on exploring the characteristics and influence of procrastination, especially psychologists and sociologists. To be specific, at first, many researches focus on exploring the reasons and personality of procrastination [10,17]. Interestingly, there also are some works devised procrastination models to depict the nature of procrastination [12,14,21]. Furthermore, many psychologists have provided practical methods to overcome procrastination [2,8,18,19]. While most of researches on procrastination are based on questionnaires by psychologists and sociologists, no work has been done by automatic methods on procrastination. As we know, the questionnaire-based measurement has several drawbacks, e.g., subjectivity and labor intensity. On the contrary, it is very promising to automatically identify user's level of procrastination by analyzing user behaviors on computer, i.e., in a complete data-driven way. Actually, we have analyzed user behaviors on computer, and found that factors of procrastination are correlated with people's usage habits on computer. This careful observation reveals that it is possible to understand and assess people's procrastination by mining computer usage log. However, to achieve this goal, there are several challenges or questions. Specifically, how to effectively bridge the gap between user behaviors on computer and psychology theories? How to automatically assess user procrastination after we have built the aforementioned relation? How to evaluate the effectiveness of the proposed procrastination assessment model?

To address the challenges mentioned above, in this paper, we provide a focused study on assessing user procrastination by mining computer usage log. Along this line, we first make a analysis on user computer logs to observe whether users have different time-series patterns. Specifically, while only computer program records in the log could not provide valuable information to explore user behaviors, to understand user behaviors, we label each recorded computer program with a class, e.g., office or media software. Based on these labeled programs, we can acquire user's behavioral patterns over time, which provides a basis for procrastination exploration. As user behaviors on computer are time-series, we propose a simple procrastination assessment model based on the Markov Chain to evaluate user procrastination. To comprehensively understand the procrastination and explore specified reasons of procrastination, we define and extract some features from the computer usage log based on psychology theories [2,12,17], which successfully bridges the gap between user behaviors on computer and psychology theories. With these extracted features, we devise a sophisticated procrastination assessment model by combining the algorithms of GBDT and CLTree, which can automatically assess user's procrastination. For precisely evaluating the effectiveness of proposed assessment model, we conduct extensive experiments with a real-world dataset and procrastination questionnaires with 115 volunteers. Experimental results are consistent with psychological findings and clearly demonstrate the accuracy of our proposed sophisticated model.

2 Preliminaries

In this section, we first illustrate the nature of procrastination, and then describe the formulation of the procrastination identification problem.

2.1 Procrastination Illustration

In the field of procrastination, Piers Steel, one of the world premier authority on the science of motivation and procrastination, points out that *"Procrastination is to voluntarily delay an intended course of action despite expecting to be worse off for the delay"* [17]. And according to the study [18] of Piers Steel, many factors of users and the task are correlated with procrastination, such as willpower, postponement, expectation and self-value, sensation seeking and so on. To understand the correlation between these factors and procrastination explicitly, we take willpower as an example:

Willpower: The willpower of users is associated with the degree of procrastination. The stronger willpower users have, the milder procrastinator users will be.

2.2 Problem Statement

As we know, procrastination becomes more pervasive among people resulting from the growing usage of Internet and computer for work. Intuitively, procrastination is correlated with users' computer behavioral logs, e.g., less working time each day implies higher degree of procrastination. In Fig. 1, we show two

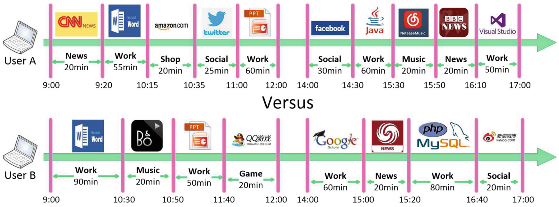


Fig. 1. Examples of two users' daily sequential behaviors.

users' sequential behaviors on computer. The green axis at the bottom indicates daily time, where the softwares are started and ended based on the associated time. According to Fig. 1, we find that the average working time (56.25 min) of user A is smaller than that of user B (70 min), which shows that user A has lower level of concentration compared with user B. What's more, user A tends to finish tasks in the last minute, while user B tends to finish jobs in advance and is better at managing time. These findings all demonstrate that user A more likely procrastinates on tasks than user B. From this example, we can see that it is possible to assess user procrastination by mining computer usage log.

Formally, given L levels of procrastination degree in which a user should belong to, we wish to capture the accurate level $l_u \in L$ of user u, so the user u can accurately know her procrastination degree and take actions to mitigate the procrastination early. Although we have no prior information on procrastination of users, we can divide users into the different groups in terms of their features F extracted from their computer logs (illustrated in Sect. 4.1) with unsupervised methods. After we obtain each user's group, we can manually label the procrastination level of each group based on the center of the group. And then we acquire every user's procrastination level. Our main assumption in this task is that users with same level of procrastination have similar usage habits on computer.

2.3 Data Collection and Preprocessing

Data Collection. We collected two data sets: users in first dataset are without procrastination levels; users in the second dataset are labeled with procrastination levels.

Unlabeled Dataset. This dataset has 1000 anonymous users' four-week behavioral logs on computer from the Chinese online-data website[1], and contains user behaviors on computer and demographic information:(1) User behaviors: userID, time, procedure name and software name; (2) Demographic information: userID, gender, birthday, education, job, income, province and city. And we filter out the users who have no statistical characteristics and unreasonable records in the data set. As a result, we get 979 users, 10,020 procedures and 7,873,723 records. With this data set, we can build the relation between the behavioral logs and factors of procrastination as described in Sect. 4.1 in details.

Labeled Dataset. Except for the unlabeled dataset, to validate the effectiveness of proposed identification models, we also conduct procrastination questionnaires with 115 volunteers. In psychology, to evaluate user's procrastination, conducting questionnaire on volunteers is a widely used method. For examples, Ferrari et al. [3] introduced that questionnaires are changed according to different tasks (academic procrastination or everyday procrastination). As we want to evaluate user's everyday procrastination and considering the effectiveness and popularity of GPS questionnaire [5], we ask these 115 volunteers to participate in the GPS questionnaire to measure their procrastination level. Specifically, we devise

[1] http://www.datatang.com/data/43910.

20 questions that can capture specific reasons (e.g., postponement, willpower) of volunteer's procrastination. Then, we ask volunteers to choose a score from 1 to 5 for each question. According to the data of returned questionnaires, we can directly label each volunteer's level (low, middle, high) of procrastination. Considering the validity of returned questionnaires, we collect 46 qualified questionnaires from these 115 volunteers. What's more, we also record these volunteers' computer usage log for a month as we have in the unlabeled dataset. Finally, we can use the labeled volunteers to evaluate the accuracy of proposed models for assessing user procrastination.

Softwares Labeling. The original data only has basic information about a procedure (e.g. time, procedure name and software name), without the type of software. Hence, we have crawled type of softwares from a popular software website[2] and add the type of softwares to the original data. While some softwares can not be labeled by the crawled rules, we label these softwares as "Unknown".

URLs Labeling. Through analyzing the dataset in depth, 32.31% records on the log are web browsing behaviors and occupy 22.23% computer using time. Intuitively, not all behaviors on browsing webpages are about entertainment, e.g., "ieee.com" probably means a working URL. Therefore, we construct 132 primary rules and 45 secondary rules to label the type (working or entertainment) of URLs. As a result, we can label 79.20% URLs totally, and 20.8% remaining URLs are labeled as "Unknown".

3 A Simple Model

Inspired by the conclusion of SM Moon et al. in [11] *"Similar users have similar behavioral patterns and contiguous behaviors are correlated with each other"*, we assume that users with similar degree of procrastination have similar behavioral patterns and the previous behavior has an effect on the current behavior. According to this hypothesis, we propose a simple procrastination identification model based on the Markov chain denoted as SPIMMarkovChain.

Formally, for each user u, given her sequential behaviors $\boldsymbol{B_u} = \{b_{ut}, t \in T\}$, the state-space $\boldsymbol{SP} = \{sp_s, s \in S\}$, we learn the behavioral types' transition matrix $\boldsymbol{tp}_u$ based on $\boldsymbol{B_u}$ and $\boldsymbol{SP}$ by the Markov chain. In our application, after labeling all procedures by methods introduced in Sect. 2.3, we can obtain a large number of procedures with types (working, entertainment and unknown). We make the set of procedure types as the state-space $\boldsymbol{SP} = \{sp_1 : entertainment, sp_2 : working, sp_3 : unknown\}$. As the Markov chain is straightforward, we omitted the detail of learning procedure on the Markov chain. Note that while we focus on the case of a single user u for ease of presentation, the extension to multiple users is easy.

[2] http://www.skycn.com.

By utilizing the Markov model, we acquire all users' transition matrix $\boldsymbol{TP} = \{\boldsymbol{tp}_u, u \in U\}$ that can approximately represent user's behavior patterns on computer. As mentioned above *"Users with similar degree of procrastination have similar behavioral patterns"*, we naturally cluster users according to the transition matrix. As dimensions of clustering features are not high, there are many clustering methods [1,16] applied to our clustering problem easily. We choose the K-Means [9] method to cluster users based on the elements of the transition matrix as its simplicity and effectiveness. We choose three elements: *tp(working* | *working)*, *tp(working* | *entertainment)*, *tp(working* | *unknown)* to cluster users as these elements are inversely proportional to user's procrastination. It is believed that values of above three elements are higher, the procrastination of users is lower.

4 A Sophisticated Model

In this section, we first extract ten features to quantify user procrastination in Sect. 4.1. And then we propose a sophisticated model for procrastination identification, namely Unsupervised Gradient Boosting Decision Tree (UGBDT) model, which can automatically and accurately evaluate user procrastination with extracted features. More details of the UGBDT model are introduced in Sect. 4.2.

4.1 Identification Features

As the SPIMMarkovChain model in Sect. 3 only considers user sequential patterns on computer, it can not depict specified reasons of user procrastination. For example, the SPIMMarkovChain model has identified user A as a serious procrastinator, but it can not provide specified reasons why user A is a serious procrastinator and corresponding solutions what user A should do to mitigate her procrastination. Therefore, we extract ten features from the computer usage log to accurately quantify user procrastination. More importantly, each feature, which comprehensively considers the psychological theories in [2,12,17] and the characteristics of users' behaviors on computer, reflects one factor of procrastination. Hence, we devise the following ten features, the first five features of user u are: *Age*, Age_u; *First Working Time*, fwt_u; *Total Working Time*, twt_u; *Total Entertaining Time*, tet_u; *Ratio between Work and Entertainment*, rwe_u. The last five are as follows:

Concentration Level: To assess the factor of user's willpower that is a strong index correlating to procrastination based on user's computer usage habit, we measure the average time: $atw_u = \frac{twt_u}{NumWR_u}$ spending on work procedures of user u, where twt_u is defined as above and $NumWR_u$ is the total number of working records in the log of user u. The larger the atw is, the longer concentration u has, and it shows that user has lower degree of procrastination.

Procrastination Length: Intuitively, the more time users spend on playing, the higher procrastination level users have. In terms of this observation, we devise a feature to the factor of postponement in the strength of procrastination. Given work and entertainment labels of users' records, we calculate the average playtime apt_u between two adjacent working records of user u.

Daily Behavior Entropy: As lower day regularity and higher disorder reflects lower conscientiousness, we can quantify the factor of user's conscientiousness by measuring the daily behavior entropy H_u of user u. As proved in psychological theories, higher conscientiousness leads to higher probability on procrastination. For this purpose, we use the Shannon Entropy [7,15]: $H_u = -\sum_{s=1}^{n} p_{us} \log p_{us}$ to measure the uncertainty of user's daily behaviors on computer, where n is the number of softwares user u has used and p_{us} is proportion of using time on software s.

Sensation Seeking: As a user who likes to seek new and adventurous things tends to procrastinate, we measure the sensation seeking of user u by calculating the rate rNT_u between the average using time aut_u of all softwares and the number ns_u of these softwares: $rNT_u = \frac{aut_u}{ns_u}$.

Software Relevance: When a user switches the current software to another one, it partially reflects the user's procrastination. In terms of this observation, we introduce "Software Relevance" of user u denoted as sr_u to measure the relevance among softwares. Small sr_u means that user u often switches to a irrelevant software from current task, which reflects that the user should spend much more time on going back to previous states and is likely to procrastinate tasks.

As different scales of these features, we first use the transformation $\frac{Max(f)-f}{Max(f)-Min(f)}$ or $\frac{f-Min(f)}{Max(f)-Min(f)}$ to normalize all features into the ranging [0,1], where f is the value of a feature. After having defined and normalized the above features that are covering procrastination's psychological theories, user's demographic information and characteristics of computer usage log, we utilize them to identify user procrastination.

4.2 Proposed Model

With these extracted features, we can adopt a clustering method to group users into different levels of procrastination. Although there are many researches about clustering [1,16], they have some drawbacks for our application. We want to develop a sophisticated model for procrastination identification that can effectively and accurately cluster users based on those extracted features. The gradient boosting decision tree (GBDT) in [4] is an additive regression model utilizing decision trees [13] as the weak learner. And the mode is nicely to our application and has some strengths [20]: (1) Well interpretability by adopting the decision trees over other learners; (2) Less prone to over-fitting by utilizing shallow decision trees. However, the GBDT is a supervised learning method, which can not be applied to our application directly as the datasets have no pre-assigned class labels. Luckily, [6] proposed a novel clustering method called the CLTree based on the supervised learning method decision tree. Inspired by this, we successfully transform the supervised GBDT classification model to the unsupervised GBDT clustering model via combing the GBDT and the CLTree, which can be used to identify user's procrastination degrees in terms of the extracting psychological features from user behaviors on computer.

Considering details of GBDT, two issues of GBDT should be addressed for clustering items: (1) GBDT can not utilize the decision tree on data without pre-assigned class labels; (2) There are no prior information for optimizing the next decision tree as no training data with labels. Inspired by the CLTree in [6], we nicely solve the two issues by reconstructing the decision tree on data without labels and adopting results of previous decision tree to boost subsequent decision tree.

Issue 1: *Clustering through decision tree* [6].

Liu et al. [6] proposed a CLTree model that is based on a supervised learning technique. Specifically, if the data have several clusters, points are not uniformly distributed in the entire space. Therefore, it is possible to partition the clusters by adding some uniformly distributed points (non-exist points), because within each cluster there are more original points than non-exist points. In terms of this observation, CLTree first regards each data in the dataset with a class Y, and then assumes that the data space is uniformly distributed with non-existing points with label N. By uniformly importing distributed non-existing points on the original data space, the problem of clustering original points turns to classifying original points Y and non-existing points N. In this way, we can adopt the decision tree to solve the transformed classifying problem.

Issue 2: *Unsupervised Gradient Boosting Decision Tree.*

The GBDT learning method utilizes the loss function to gradually boost the effect of next decision tree. Nevertheless, GBDT can not construct a loss function on the data without classes, which results in that GBDT is not capable of improving the effectiveness and accuracy of clusters iteratively. However, considering the solution of issue one, we can use a subset of features space to construct the first unsupervised decision tree by CLTree and obtain preliminary clusters of Y points on the first decision tree. What information can we use from results of the first decision tree? Intuitively, the points in same cluster from the first decision tree have higher probability in same cluster than those points in different clusters. In terms of this observation, we utilize this information to guide next decision tree's construction.

Formally, during current decision tree's construction, if the split makes points within previous decision tree's same cluster into different nodes, we will punish this split as it takes chaos to the node. We take the chaos as a penalty term and is named cluster entropy. By introducing the cluster entropy as a penalty term, we can obtain a adjusted information gain to select the best cut in current decision tree m as following:

$$ag(D, A) = g(D, A) + \sum_{j=1}^{J} \alpha_j \Big(- \sum_{i=1}^{n} \frac{|D_{Yj,i}^{(m-1)}|}{|D_{Yj}^{(m-1)}|} \log \frac{|D_{Yj,i}^{(m-1)}|}{|D_{Yj}^{(m-1)}|} \Big), \tag{1}$$

where α_j is the penalty parameter of cluster j, J is the number of clusters from the $(m-1)$th tree and $|D_{Yj,i}^{(m-1)}|$ is the number of Y points labeled the $(m-1)$th tree's cluster id j on current decision tree's node i. By Eq. 1, we can iteratively boost the result of the decision tree by adopting previous cluster results and solve the issue two.

For procrastination identification, we input the extracted features in Sect. 4.1 to the unsupervised GBDT model. On the above, we have successfully transformed the supervised GBDT classifying method to the unsupervised clustering method by combining the CLTree and reconstructing the GBDT. The supervised GBDT also can be used to other classification problems with good interpretability.

5 Experiments

In this section, we conduct extensive experiments on two real-world data sets for validating the effectiveness of proposed models.

5.1 Baseline Methods and Evaluation Metrics

Since user's procrastination identification is a clustering problem, we adopt KMeans as one baseline, which is a representative clustering method and widely used in practical works. What's more, our sophisticated UGBDT method refers to the design of CLTree. Spontaneously, CLTree is selected as a baseline, too.

To evaluate the proposed models comprehensively, we adopt two categories of evaluation metrics:

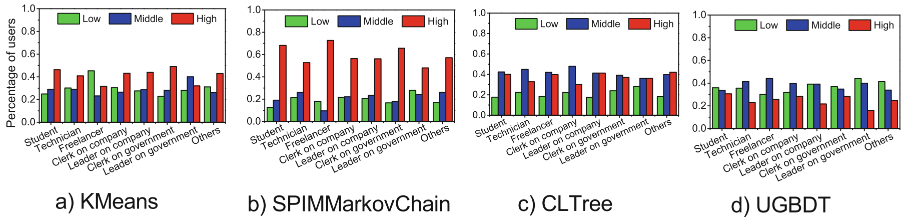
Evaluation Metrics for Psychology. As the purpose of proposed models is to assess user's level of procrastination, we must evaluate that whether the results conform to the psychological findings on procrastination. Inspired by "The procrastination is correlated with jobs and ages: the procrastination of students is the most serious and the procrastination of users tends to decrease with age" in [18], we extract two psychological findings: (1) The procrastination of students is the most serious compared with other jobs; (2) The procrastination of users tends to decrease with age. We adopt these two findings to validate whether the results assessed by proposed models conform to psychological theories.

Accuracy. As we have one dataset with labeled procrastination levels of people, we intuitively compute the accuracy of proposed models based on the labeled dataset, which can accurately evaluate proposed models on assessing people's procrastination. Particularly, larger accuracy indicates better performance of procrastination assessment.

5.2 Performance on Unlabeled Dataset

In this section, we present experimental results on unlabeled dataset introduced in Sect. 2 for validating the proposed models' performance on assessing people's procrastination. As people on the first dataset are unlabeled, it means that we do not know the actual level of people's procrastination. Although the procrastination of these unlabeled users are unknown, they have demographic information (e.g., jobs and ages) that can be utilized to validate the psychological findings on procrastination:

Distributions of Users on Jobs. To validate the finding "The procrastination of students is the most serious compared with other jobs", we exhibit users' distributions with different procrastination levels on jobs as shown in Fig. 2. According to Fig. 2, we can draw several implications: (1) The result of UGBDT shows that the procrastination of students (30.6%) is most serious than other jobs, while the job with most serious procrastination of KMeans, MarkovChain and CLTree are "Clerk on government", "Freelancer" and "Others", respectively; (2) The result of UGBDT also shows that the procrastination of leaders on government and company is lower than clerk, while other baselines can not validate this psychological finding.



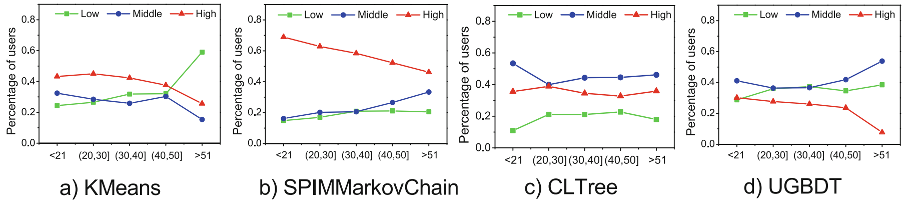
a) KMeans b) SPIMMarkovChain c) CLTree d) UGBDT

Fig. 2. Distributions of users on jobs with models.

Distributions of Users on Ages. Next, in order to validate the finding "The procrastination of users tends to decrease with age", we exhibit users' distributions with different procrastination levels on ages in Fig. 3. According to Fig. 3, we can draw several implications: (1) The results of UGBDT and SPIMMarkovChain show that the degree of procrastination obviously decreases as the age of users increases, while the trend of decrease based on KMeans and CLTree are not obvious compared with UGBDT and SPIMMarkovChain; (2) The users' distributions of different levels of procrastination by UGBDT and CLTree are more truthful than other two baselines, because users with high procrastination grouping by SPIMMarkovChain and KMeans are much more than users with middle or low procrastination. In terms of these two findings, we find that only UGBDT can both effectively capture the trend of users' procrastination on ages and actually group users into different levels of procrastination.

5.3 Performance on Labeled Dataset

Except for the experiments on unlabeled dataset, we also conducted experiments on labeled dataset to evaluate the effectiveness of proposed models. Similar to



a) KMeans b) SPIMMarkovChain c) CLTree d) UGBDT

Fig. 3. Distributions of users on ages with models.

Sect. 5.2, we also adopt KMeans and CLTree as baselines to compare the accuracy of proposed models (SPIMMarkovChain and UGBDT). For users on labeled dataset, we can group them into three clusters by these methods, and manually label the three clusters with corresponding procrastination levels by clustering features. Then, we compare each user's procrastination level by clustering methods with the labeled levels on the dataset. Finally, we can obtain each method's accuracy on evaluating user's procrastination level as: *KMeans, 52.2%; SPIMMarkovChain, 43.5%; CLTree, 50.0%; UGBDT, 60.9%.* These results demonstrate that the UGBDT has the largest accuracy (60.9%) compared with other methods, which means that UGBDT is the best method on identifying user's procrastination. Particularly, the result (52.2%) of KMeans based on the extracted features outperforms better than SPIMMarkovChain (43.5%) based on behavioral patterns, which demonstrates the effectiveness of extracted features.

5.4 Exploration of Procrastination Reasons

Compared with the simple model in Sect. 3, the sophisticated UGBDT model can reveal factors that may cause user's procrastination based on extracted features. In this section, we show a case study on three questionnaire users with different procrastination level to evaluate the capacity on depicting specified reasons of user procrastination.

We show the captured procrastination reasons of these three users by the sophisticated model. We exhibit the extracted features of these three selected users in Fig. 4. According to Fig. 4, we can find that most of extracted features

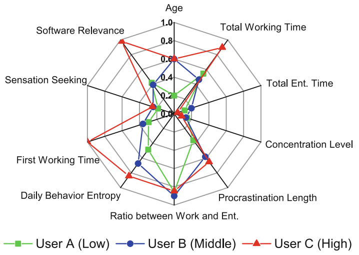


Fig. 4. A case study on three users.

(e.g., "First Working Time" and "Procrastination Length") of the user with higher procrastination are generally greater than the user with lower procrastination, and that the reasons causing the severe procrastination of user C are severe postponement and lower expectation and self-value, which is conforming to the observation of user questionnaires. If user C wants to mitigate her procrastination, she should improve the action and power of focus. Considering computer usage behaviors, user C should improve the working time and focus on current tasks, avoiding disturbs of other things.

6 The Conclusion

In this paper, we provided a focused study on understanding and assessing people's procrastination by mining users' behavioral log on computer. As user's behaviors on computer are time-series, we first proposed a simple procrastination assessment model based on the Markov Chain, which could directly capture user's behavioral patterns and evaluate procrastination. However, this simple model could not depict reasons of procrastination. To explore possible reasons of user procrastination, we then extracted some features from computer usage log to quantify procrastination, which successfully bridges the gap between user behaviors on computer and psychological theories. With extracted features, we devised a sophisticated procrastination assessment model by combining the algorithms of GBDT and CLTree, which can accurately assess the level of user procrastination. More importantly, the sophisticated model can reveal factors that may cause people's procrastination. To validate the effectiveness of proposed models, we conducted extensive experiments with unlabeled dataset and procrastination questionnaires with 115 volunteers, and the experimental results clearly demonstrate the effectiveness of our proposed sophisticated model. To the best of our knowledge, this work is the first to automatically assess people's procrastination based on computer usage log and successfully transform the supervised GBDT classification model to the unsupervised GBDT clustering model. Also, this work could provide valuable insights for psychology/behavior researchers to further explore procrastination.

Acknowledgements. This research was partially supported by grants from the National Natural Science Foundation of China (Grants No. 61727809, U1605251, 61672483, 61602234 and 61572032), and the Science Foundation of Ministry of Education of China & China Mobile (No. MCM20170507).

References

1. Berkhin, P.: A survey of clustering data mining techniques. In: Kogan, J., Nicholas, C., Teboulle, M. (eds.) Grouping Multidimensional Data, pp. 25–71. Springer, Heidelberg (2006). https://doi.org/10.1007/3-540-28349-8_2
2. Burka, J.B., Yuen, L.M.: Procrastination: Why You Do it, What to Do About it Now. Da Capo Press, Cambridge (2008)

3. Ferrari, J.R., Johnson, J.L., McCown, W.G.: Assessment of academic and everyday procrastination. Procrastination and Task Avoidance. The Springer Series in Social Clinical Psychology, pp. 47–70. Springer, Boston (1995). https://doi.org/10.1007/978-1-4899-0227-6_3
4. Friedman, J.H.: Stochastic gradient boosting. Comput. Stat. Data Anal. **38**(4), 367–378 (2002)
5. Lay, C.H.: At last, my research article on procrastination. J. Res. pers. **20**(4), 474–495 (1986)
6. Liu, B., Xia, Y., Yu, P.S.: Clustering through decision tree construction. In: Proceedings of the Ninth International Conference on Information and Knowledge Management, pp. 20–29. ACM (2000)
7. Liu, Q., Chen, E., Xiong, H., Ge, Y., Li, Z., Wu, X.: A cocktail approach for travel package recommendation. Knowl. Data Eng. IEEE Trans. **26**(2), 278–293 (2014)
8. Liu, Q., Zeng, X., Zhu, H., Chen, E., Xiong, H., Xie, X., et al.: Mining indecisiveness in customer behaviors. In: IEEE International Conference on 2015 Data Mining (ICDM), pp. 281–290. IEEE (2015)
9. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Oakland, vol. 1, pp. 281–297 (1967)
10. McCown, W., Johnson, J., Petzel, T.: Procrastination, a principal components analysis. Pers. Individ. Differ. **10**(2), 197–202 (1989)
11. Moon, S.M., Illingworth, A.J.: Exploring the dynamic nature of procrastination: a latent growth curve analysis of academic procrastination. Pers. Individ. Differ. **38**(2), 297–309 (2005)
12. Procee, R., Kamphorst, B.A., Wissen, A.v., Meyer, J.J.C.: An agent-based model of procrastination. In: ECAI 2014–21st European Conference on Artificial Intelligence, 18–22 August 2014, Prague, Czech Republic-Including Prestigious Applications of Intelligent Systems (PAIS 2014), vol. 263, pp. 747–752. IOS Press (2014)
13. Quinlan, J.R.: Simplifying decision trees. Int. J. Man Mach. Stud. **27**(3), 221–234 (1987)
14. Ross, D.: Economic models of procrastination. Legal Ethics (2010)
15. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE Mob. Comput. Commun. Rev. **5**(1), 3–55 (2001)
16. Sim, K., Gopalkrishnan, V., Zimek, A., Cong, G.: A survey on enhanced subspace clustering. Data Min. Knowl. Discov. **26**(2), 332–397 (2013)
17. Steel, P.: The nature of procrastination: a meta-analytic and theoretical review of quintessential self-regulatory failure. Psychol. Bull. **133**(1), 65 (2007)
18. Steel, P.: The procrastination equation: how to stop putting things off and start getting stuff done. Random House Canada (2010)
19. Wu, L., Liu, Q., Chen, E., Xie, X., Tan, C.: Product adoption rate prediction: a multi-factor view. In: Proceedings of the 2015 SIAM International Conference on Data Mining, pp. 154–162. SIAM (2015)
20. Ye, J., Chow, J.H., Chen, J., Zheng, Z.: Stochastic gradient boosted distributed decision trees. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 2061–2064. ACM (2009)
21. Zhu, Y., Zhu, H., Liu, Q., Chen, E., Li, H., Zhao, H.: Exploring the procrastination of college students: a data-driven behavioral perspective. In: Navathe, S., Wu, W., Shekhar, S., Du, X., Wang, X., Xiong, H. (eds.) International Conference on Database Systems for Advanced Applications, vol. 9642, pp. 258–273. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-319-32025-0_17