Exploring the Procrastination of College Students: A Data-Driven Behavioral Perspective

Yan Zhu[1], Hengshu Zhu[2], Qi Liu[1], Enhong Chen[1](✉), Hong Li[3], and Hongke Zhao[1]

[1] University of Science and Technology of China, Hefei, China
{zhuyan90,zhhk}@mail.ustc.edu.cn, {qiliuql,cheneh}@ustc.edu.cn
[2] Baidu Research-Big Data Lab, Beijing, China
zhuhengshu@baidu.com
[3] Hefei University, Hefei, China
xiaoke_93@126.com

Abstract. Procrastination refers to the practice of putting off impending tasks due to the habitual carelessness or laziness. The understanding of procrastination plays an important role in educational psychology, which can help track and evaluate the comprehensive quality of students. However, traditional methods for procrastination analysis largely rely on the knowledge and experiences from domain experts. Fortunately, with the rapid development of college information systems, a large amount of student behavior records are captured, which enables us to analyze the behaviors of students in a quantitative way. To this end, in this paper, we provide a data-driven study from a behavioral perspective to understand the procrastination of college students. Specifically, we propose an unsupervised approach to quantitatively estimate the procrastination level of students by the analysis of their borrowing records in library. Along this line, we first propose a naive Reading-Procrastination (naive RP) model, which considers the behavioral similarity between students for procrastination discovery. Furthermore, to improve the discovery performance, we develop a dynamic Reading-Procrastination (dynamic RP) model by integrating more comprehensive characteristics of student behaviors, such as semester-awareness and month-regularity. Finally, we conduct extensive experiments on several real-world data sets. The experimental results clearly demonstrate the effectiveness of our approach, and verify several key findings from psychological fields.

1 Introduction

Procrastination refers to the practice of putting off impending tasks to a later time, sometimes to the "last minute" before a deadline, which is usually due to the habitual carelessness or laziness [18]. As a matter of fact, exploring procrastination plays an important role in educational psychology, which can help track and evaluate the comprehensive quality of students.

S.B. Navathe et al. (Eds.): DASFAA 2016, Part I, LNCS 9642, pp. 258–273, 2016.
DOI: 10.1007/978-3-319-32025-0_17

However, psychological fields often focus on analyzing the causes [11,14] and effects [2,6] of procrastination, few efforts have been devoted to quantitatively discovering the procrastination of students. Meanwhile, traditional methods for procrastination analysis largely rely on the knowledge and manual labor of domain experts, such as questionnaire and survey. Such self-reported approach has been found that it has only a moderate correlation with observed procrastination [13]. Therefore, it is appealing to develop an approach to automatically uncover the procrastination behavior, which is still under-addressed.

Fortunately, thanks to the rapid development of college information system, a large amount of behavior records of students are captured, which opens a better venue for analyzing students habits. To this end, we introduce a data-driven study from a behavioral perspective to explore the procrastination of college students. Specifically, we propose an unsupervised approach to quantitatively estimate the procrastination level of students through the analysis of their borrowing records in college library. Particularly, based on the research findings in psychological studies, we assume that procrastination is a latent factor which may affect the hold time of borrowed books. Therefore, instead of directly mining procrastination phenomenon, we propose to comprehensively model the borrowing behaviors of students and probe the procrastination through predicting the hold time of borrowed books in library. Along this line, we propose a naive Reading-Procrastination (naive RP) model, which takes consideration of the behavioral similarity between students for procrastination discovery. To improve the discovery performance, we develop a dynamic Reading-Procrastination (dynamic RP) model by integrating more comprehensive characteristics of student behaviors, such as semester-awareness and month-regularity. A unique characteristic of the dynamic RP model is that it can depict the procrastination behavior in a probabilistic and empirical Bayesian perspective. Finally, extensive experiments are carried out on real world data sets collected from a Chinese college. The experimental results demonstrate the effectiveness of our approach, and verify several key findings from psychological fields.

2 Preliminaries

In this section, we first introduce the details of our real world data. Then, we present the basis of procrastination assumption, and finally describe the formulation of the procrastination discovery problem.

2.1 Data Description

The data set used in our study is the library borrowing records provided by a Chinese four-year university. The snapshot of the data set is shown as Table 1. In addition, the data set also includes the profile information of students such as grade, major and sex. In particular, since library usually has limited volumes of each book, students must comply with some restricted borrowing rules made by college library. First, each student can hold at most a limited number of

Table 1. A snapshot of library borrowing records

User id	Book id	Borrowing time	Due time	Return time
U_1	B_5	2010-04-01 09:05:20	2010-07-31 23:59:59	2010-07-12 16:12:16
U_2	B_2	2010-04-01 09:08:22	2010-07-31 23:59:59	2010-05-02 19:25:08
U_2	B_6	2010-04-01 09:08:45	2010-07-31 23:59:59	2010-06-24 10:05:36
...	...	...	...	...
U_3	B_1	2010-04-13 14:16:17	2010-08-12 23:59:59	2010-05-15 12:01:03
U_4	B_2	2010-04-13 14:16:55	2010-08-12 23:59:59	2010-06-44 11:45:56

(e.g. six) books in total. That is to say, students who have already had six books at hand must return at least one book if they want to borrow another one. Second, each book can be held for a limited number of (e.g. 120) days in fairness to all of students. To explore the procrastination behavior of students, we choose the records of students entering school from 2006 to 2010 as study candidates, including 17,531 students and 107,818 books with 1,007,406 borrowing records.

2.2 Procrastination Assumption

A study of academic procrastination reveals that 46 % of subjects reported they nearly always or always procrastinate on writing a paper, while 27.6 % procrastinate on studying for examinations [10]. Furthermore, it is estimated that 80 %–95 % of college students engage in procrastination, approximately 75 % consider themselves procrastinators [11]. Therefore, the procrastination of students is very likely to influence their behaviors in library, and the extent of which is determined by the level of procrastination. Based on the above, here we assume that procrastination is a latent factor which may affect the hold time of borrowed books. Therefore, such factor can be learned through comprehensively modeling the borrowing behaviors of students.

2.3 Problem Formulation

Since the procrastination of students is not an explicit variable, we cannot directly estimate the value of procrastination through supervised approach. Therefore, we need to find some quantitative observations and seek the relationship between the procrastination and them. Based on our procrastination assumption, we can regard procrastination as a latent factor that determines the hold time of borrowed books. Through the modeling of hold time, we can learn the procrastination value for each student. If the hold time of borrowed books can be predicted accurately, it can be verified that procrastination surely is an attribute of college students and undoubtedly makes effect on their hold time of borrowed books. For simplicity, we assume the procrastination value of each student is invariable over the four years in college.

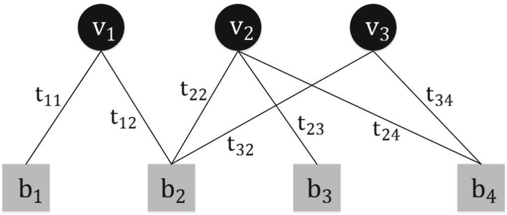


Fig. 1. Example of the B-S bipartite network.

Formally, we define e_u as the procrastination value of student u. The hold time of the book i borrowed by students u at time t is represented by $h_{u,i}(t)$. Our task is to model the causal relationship $e_u \to \{h_{u,i}(t)\}$ from e_u to a set of borrowing records $\{h_{u,i}(t)\}$ of student u. In this way, the procrastination value e_u can be obtained as a result of the hold time modeling.

3 Naive RP Model

Based on procrastination assumption, we can estimate the hold time of borrowed book $h_{u,i}(t)$ through procrastination value e_u. However, intuitively, $h_{u,i}(t)$ is not completely determined by e_u, since different books have different *required reading time*. Therefore, we should first clarify the *required reading time* $r_{u,i}(t)$ for student u and the borrowed book i at time t. The *required reading time* is not equal to the actual hold time of borrowed book, but is just an estimation of reading time that student u may spend in reading book i due to different reading speed and different focused content. Suppose this *required reading time* has been obtained, we can make estimation of hold time $h_{u,i}(t)$ by $r_{u,i}(t) + e_u$. By minimizing the error function, we can learn the procrastination value e_u for all of students. The error function is represented as:

$$\min_{\{e_u|u \in U\}} \sum_{(u,i,t) \in R} (h_{u,i}(t) - r_{u,i}(t) - e_u)^2 + \sum_{u \in U} \lambda e_u^2, \tag{1}$$

where R is the set of borrowing records and U is the set of students. In order to make a tradeoff between the magnitude of *required reading time* and procrastination factor, we add a regularization term λe_u^2 into the above error function.

3.1 Required Reading Time Estimation

The *required reading time* $r_{u,i}(t)$ is an important component of the hold time $h_{u,i}(t)$ and its estimation also has impact on precise estimation of procrastination level of each student. However, less information about the student can be used to describe this *required reading time*. As for book i, it is impractical to analyze its content for reading time estimation due to the lack of electronic data. Therefore, we propose to leverage the neighborhood approach for estimation and define $r_{u,i}(t)$ through borrowing history of book i:

$$r_{u,i}(t) = \sum_{v \in L(i)} w_i(u,v) h_{v,i}, \tag{2}$$

where $L(i)$ is the set of students who borrowed book i in borrowing history. $h_{v,i}$ is the vth student's hold time for book i. Thus, it raises an issue that how to determine the value of weight $w_i(u,v)$ in order to estimate the *required reading time*. Directly, we can define $w_i(u,v)$ by means of similarity $sim_i(u,v)$ between student u and his neighbors. To this end, we build a book-student (B-S) bipartite network $G = \{V, B, T\}$ as shown in Fig. 1, where $V = \{v_1, ..., v_{|V|}\}$ denotes the set of students in library history, and $B = \{b_1, ..., b_{|B|}\}$ denotes the set of borrowed books. $T = \{t_{vi}\}$ is the edge set, where t_{vi} denotes the time student v borrowed book i previously. For convenience of calculating the similarity between students, in this paper, we let $t_{vi} = (t^s_{vi}, t^m_{vi}, t^d_{vi})$, where t^s_{vi}, t^m_{vi} and t^d_{vi} represents the semester, month and day when student v borrowed book i. Thus, we can define the similarity $sim^1_i(u,v)$ as:

$$sim^1_i(u,v) = \frac{e^{I(q_u = q_v)}}{1 + |t^m_{ui} - t^m_{vi}|}, \tag{3}$$

where $I(q_u = q_v)$ is the indicator function, q_u and q_v denote the major of student u and v respectively. It is sound that students who majored in same field might focus on the same content of the book. Moreover, the more adjacent month students borrow this book at one year, the more likely they sign up for the same course and the more similar knowledge they need to learn from this book.

However, to depict the similarities of students, these observational similarities are not enough when estimating the *required reading time* of borrowed books. Thus, we need to define the similarities from different point of view. As mentioned in preliminaries section, students must comply with the borrowing volume number constraint. So students who frequently borrow books relatively have less hold time of borrowed books. In this way, we incorporate this borrowing frequency factor into the similarity definition. In this paper, we only consider the borrowing frequency in each semester. Specifically, we first seek out edge set $T^s_v = \{t_{vj} | t^s_{vj} = t^s_{vi}\}$, which is the set of dates student v had borrowing actions in semester t^s_{vi}. Second, we divide semester into several timestamps and count the borrowing number in each timestamp utilizing t^m_{vi} and t^d_{vi} in set T^s_v. Then, we define a vector $\overrightarrow{n_{v,i}}$ to represent the frequency, in which each entry corresponds to the above counting borrowing number in each timestamp. After obtaining vector $\overrightarrow{n_{u,i}}$ and $\overrightarrow{n_{v,i}}$, we can compute similarity conveniently. Motivated by Tanimoto similarity coefficient [15], which is a generalized Jaccard similarity coefficient, and is defined as:

$$T_s(\overrightarrow{x}, \overrightarrow{y}) = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{|\overrightarrow{x}|^2 + |\overrightarrow{y}|^2 - \overrightarrow{x} \cdot \overrightarrow{y}} = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{|\overrightarrow{x} - \overrightarrow{y}|^2 + \overrightarrow{x} \cdot \overrightarrow{y}}, \tag{4}$$

we define the similarity $sim^2_i(u,v)$ as:

$$sim^2_i(u,v) = \frac{\cos < \overrightarrow{n_{u,i}}, \overrightarrow{n_{v,i}} >}{|\overrightarrow{n_{u,i}} - \overrightarrow{n_{v,i}}| + \cos < \overrightarrow{n_{u,i}}, \overrightarrow{n_{v,i}} >}. \tag{5}$$

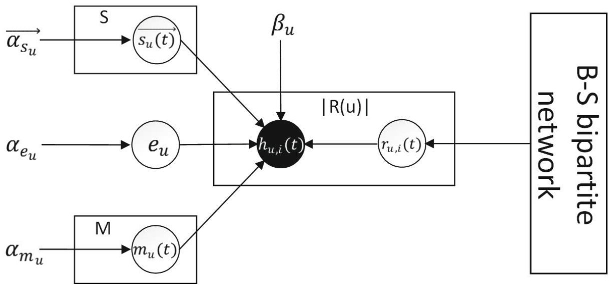


Fig. 2. The graphical representation of dynamic RP model.

In Eq. 5, we consider not only the frequency but also the magnitude in similarity computation. For example, if $\overrightarrow{n_{u,i}} = (2,4)$, and $\overrightarrow{n_{v1,i}} = (1,4)$, $\overrightarrow{n_{v2,i}} = (2,2)$, $\overrightarrow{n_{v3,i}} = (2,4)$, $\overrightarrow{n_{v4,i}} = (4,8)$, then $sim_i^2(u,v1) = 0.4940$, $sim_i^2(u,v2) = 0.3217$, $sim_i^2(u,v3) = 1$ and $sim_i^2(u,v4) = 0.1827$. We can find that $sim_i^2(u,v4)$ is the smallest although $\overrightarrow{n_{v4,i}}$ is proportional to $\overrightarrow{n_{u,i}}$. Taking account of $sim_i^1(u,v)$ and $sim_i^2(u,v)$, we represent $w_i(u,v)$ as:

$$w_i(u,v) = \frac{sim_i^1(u,v) + sim_i^2(u,v)}{\sum_{v=1}^{L_i} sim_i^1(u,v) + sim_i^2(u,v)}. \tag{6}$$

3.2 Limitation Discussion

Having estimated the *required reading time* $r_{u,i}(t)$ and learned the procrastination value e_u, we can also predict the hold time of future borrowed books for student u. However, the naive RP model assumes that all students share the same regularization coefficient λ, which may be unsuitable for all students. Besides, other factors except procrastination may also affect the hold time of borrowed books, which have impact on the precision of procrastination estimation. To this end, a graphical model named dynamic RP model is developed, where the procrastination value e_u of every student is determined all by its own. In the next section, we will illustrate how dynamic RP model can capture these factors as well as can control estimation accuracy automatically.

4 Dynamic RP Model

In this section, we analyze some potential factors that dynamically influence the hold time of borrowed books. To depict these dynamic characteristics, we develop our dynamic RP model from probabilistic point of view, where the parameters estimations are under the framework of empirical Bayes.

4.1 Hold Time Component Elements

The dynamic RP model is developed on the basis of naive RP model and is shown in Fig. 2, where unshaded variables indicate latent factors that determine the hold time. In Fig. 2, $h_{u,i}(t)$ and $r_{u,i}(t)$ stand for the hold time and the *required reading time* of book i borrowed by students u at time t, respectively. e_u represents the above-mentioned procrastination value and we assume it follows a zero-mean Gaussian distribution with a parameter α_{e_u} for student u:

$$P(e_u|\alpha_{e_u}) = \mathcal{N}(e_u|0, \alpha_{e_u}^{-1}). \tag{7}$$

Naturally, besides the *required reading time*, procrastination is not the only factor that affects the hold time of borrowed books. In college, students making borrowing actions at different time of semester usually have different type of borrowing patterns. Some students tend to borrow books for learning guidance at the beginning of each semester, while others prefer to borrow at the end of each semester just for final examinations. Therefore, the hold time of their borrowed books are influenced by the specific time of each semester. Besides, course teachers assign tasks with various levels in each semester, which also has impact on the hold time of borrowed books. To capture this point, we specify $\overrightarrow{s_u(t)}$ to describe these factors, as shown in Fig. 2. Specifically, we select several time points in each semester as *kernel points*. The more close the borrowing actions take place to a specific kernel, the greater degree the hold time falls into that pattern. This factor $f_s(t)$ is defined formally as:

$$f_s(t) = \frac{\sum_{k=1}^{K} e^{-|t-t_k^0|} \cdot s_{u,k}(t)}{\sum_{k=1}^{K} e^{-|t-t_k^0|}}, \tag{8}$$

where K is the number of *kernel points*, $s_{u,k}(t)$ is the kth entry in $\overrightarrow{s_u(t)}$, t_k^0 stands for the date of kth *kernel point*. Therefore, different dates in the same semester share the same value of $\overrightarrow{s_u(t)}$, and there are totally S semesters. Again, we choose a form of zero-mean Gaussian for $s_{u,k}$ with a parameter $\alpha_{s_{u,k}}$ for student u, in which $\alpha_{s_{u,k}}$ is the kth entry in $\overrightarrow{\alpha_{s_u}}$, and

$$P(s_{u,k}(t)|\alpha_{s_{u,k}}) = \mathcal{N}(s_{u,k}(t)|0, \alpha_{s_{u,k}}^{-1}). \tag{9}$$

Furthermore, it is common that students may be free this month and become busy next month due to various reasons. These unpredictable factors come out of some stochastic events, which can alter students' reading schedule and then determine the hold time of their borrowed books. With this consideration, we choose the variable m_u to denote those month regularity factors. As shown in Fig. 2, the variable $m_u(t)$ is also shared by different dates in the same month and there are M months in total. A zero-mean Gaussian distribution with a parameter α_{m_u} is also available for m_u.

$$P(m_u(t)|\alpha_{m_u}) = \mathcal{N}(m_u(t)|0, \alpha_{m_u}^{-1}). \tag{10}$$

These above discussed factors are the supplementary component elements of the hold time of borrowed books. Adding these elements can not only modeling the hold time of borrowed books more comprehensively, but also make the estimation of procrastination value more precise. In practice, if students have borrowed very few books previously, it is better to use naive RP model since these dynamic factors can not be captured by dynamic RP model adequately. In contrast, when there are enough borrowing records, we choose dynamic RP model to depict those dynamic characteristics of each student. The model training process is presented as follow.

4.2 Empirical Bayes Framework for Estimation Accuracy Control

Based on the discussion above, we prepare to train dynamic RP model with respect to the procrastination value e_u and the above-mentioned variables. For convenience, we define $\Theta_u = \{e_u, \overrightarrow{s_u(t)}, m_u(t)\}$ and $\Lambda_u = \{\alpha_{e_u}, \overrightarrow{\alpha_{s_u}}, \alpha_{m_u}\}$. The likelihood function is given by:

$$P(\overrightarrow{h_u}|r_{u,i}(t), \Theta_u, \beta_u) = \prod_{(u,i,t)\in R(u)} \mathcal{N}(h_{u,i}(t)|p_{ui}(t), \beta_u^{-1}), \tag{11}$$

$$p_{u,i}(t) = r_{u,i}(t) + e_u + \frac{\sum_{k=1}^{K} e^{-|t-t_k^0|} \cdot s_{u,k}(t)}{\sum_{k=1}^{K} e^{-|t-t_k^0|}} + m_u(t), \tag{12}$$

where $\overrightarrow{h_u}$ is the vector of $h_{u,i}(t)$ in students borrowing set $R(u)$. In Eq. 11, we also assume $h_{u,i}(t)$ is characterized by a Gaussian distribution with a parameter β_u for student u. Due to the conjugate property, the posterior distributions of Θ_u are also Gaussian and they are represented by:

$$\begin{aligned} P(e_u|\overrightarrow{h_u}, \Theta_u, \alpha_{e_u}) &= \mathcal{N}(e_u|\mu_{e_u}, \lambda_{e_u}), \\ P(s_{u,k}(t)|\overrightarrow{h_u}, \Theta_u, \alpha_{s_{u,k}}) &= \mathcal{N}(s_{u,k}(t)|\mu_{s_{u,k}(t)}, \lambda_{s_{u,k}}), \\ P(m_u(t)|\overrightarrow{h_u}, \Theta_u, \alpha_{m_u}) &= \mathcal{N}(m_u(t)|\mu_{m_u(t)}, \lambda_{m_u}), \end{aligned} \tag{13}$$

where the means and variances of Θ_u are calculated as:

$$\begin{aligned} \mu_{e_u} &= \beta_u \lambda_{e_u}^{-1} \sum_{(u,i,t)\in R(u)} (h_{u,i}(t) - p_{u,i}(t) + e_u), \\ \lambda_{e_u} &= \alpha_{e_u} + \beta_u |R(u)|, \\ \mu_{s_{u,k}(t)} &= \beta_u \lambda_{s_{u,k}}^{-1} \sum_{(u,i,t)\in R_s(u,t)} W_k(t)(h_{u,i} - p_{u,i} + W_k(t) s_{u,k}(t)), \\ \lambda_{s_{u,k}} &= \alpha_{s_{u,k}} + \beta_u \sum_{(u,i,t)\in R_s(u,t)} W_k(t)^2, \\ \mu_{m_u(t)} &= \beta_u \lambda_{m_u}^{-1} \sum_{(u,i,t)\in R_m(u,t)} (h_{u,i}(t) - p_{u,i}(t) + m_u(t)), \\ \lambda_{m_u} &= \alpha_{m_u} + \beta_u |R_m(u,t)|, \end{aligned}$$

where $R(u)$, $R_s(u,t)$ and $R_m(u,t)$ are the set of borrowing records for student u generated in all four years, in the semester of t and in the month of t, respectively. $W_k(t)$ is given by:

$$W_k(t) = \frac{e^{-|t-t_k^0|}}{\sum_{j=1}^{K} e^{-|t-t_j^0|}}.$$

Suppose the priors Λ_u have been obtained, which are inferred from the data and will be discussed later. By maximizing the posterior distributions of Θ_u, the procrastination value of student u and other dynamic variables can be estimated by the means of Eq. 13. With these obtaining variables, we can also make prediction for new instance of hold time of borrowed book. The predictive distribution of $h_{u,\hat{i}}(t)$ takes the form of:

$$\begin{aligned} P(h_{u,\hat{i}}(t)|\overrightarrow{h_u}, r_{u,i}(t), \Lambda_u, \beta_u) = \int P(h_{u,\hat{i}}(t)|r_{u,i}(t), \Theta_u, \beta_u) \\ \times P(e_u|\overrightarrow{h_u}, \Theta_u, \alpha_{e_u}) P(\overrightarrow{s_u(t)}|\overrightarrow{h_u}, \Theta_u, \overrightarrow{\alpha_{s_u}}) P(m_u(t)|\overrightarrow{h_u}, \Theta_u, \alpha_{m_u}) \mathrm{d}\Theta_u. \end{aligned} \tag{14}$$

As stated above, both the conditional distribution of $h_{u,i}(t)$ and the distributions of Θ_u are all Gaussian distribution, which makes it possible to derive the closed form solution. By means of integral operations, the mean of predictive distribution is obtained using the result that substituting the means of posterior distributions of Θ_u into Eq. 12 simply.

Now we discuss how to get appropriate priors Λ_u from the data. The marginal likelihood function is represented by:

$$\begin{aligned} P(\overrightarrow{h_u}|r_{u,i}(t), \Lambda_u, \beta_u) = \int P(\overrightarrow{h_u}|r_{u,i}(t), \Theta_u, \beta_u) \\ \times P(e_u|\alpha_{e_u}) P(\overrightarrow{s_u(t)}|\overrightarrow{\alpha_{s_u}}) P(m_u(t)|\alpha_{m_u}) \mathrm{d}\Theta_u, \end{aligned} \tag{15}$$

where $\overrightarrow{\alpha_{s_u}}$ and $\overrightarrow{s_u(t)}$ are K-dimensional vectors that each entry of it corresponds to $\alpha_{s_{u,k}}$ and $s_{u,k}(t)$, respectively. Fortunately, all terms in Eq. 15 are Gaussian, which comes out a closed form when integrating over parameters Θ_u. Thus, we obtain the following results by maximizing Eq. 15.

$$\begin{aligned} \beta_u &= \frac{|R(u)| - \gamma_{e_u} - \sum_{k=1}^{K} \gamma_{s_{u,k}} - \gamma_{m_u}}{\sum_{(u,i,t)\in R(u)} (h_{u,i}(t) - p_{ui}(t))^2}, \\ \alpha_{e_u}^{-1} &= e_u^2 \gamma_{e_u}^{-1}, \qquad \gamma_{e_u} = \frac{\lambda_{e_u} - \alpha_{e_u}}{\lambda_{e_u}}, \\ \alpha_{s_{u,k}}^{-1} &= \frac{\gamma_{s_{u,k}}^{-1}}{|R_s(u,t)|} \sum_{(u,i,t)\in R_s(u,t)} s_{u,k}(t)^2, \qquad \gamma_{s_{u,k}} = \frac{\lambda_{s_{u,k}} - \alpha_{s_{u,k}}}{\lambda_{s_{u,k}}}, \\ \alpha_{m_u}^{-1} &= \frac{\gamma_{m_u}^{-1}}{|R_m(u,t)|} \sum_{(u,i,t)\in R_m(u,t)} m_u(t)^2, \qquad \gamma_{m_u} = \frac{\lambda_{m_u} - \alpha_{m_u}}{\lambda_{m_u}}. \end{aligned}$$

Afterwards, we can substitute above results into Eq. 13 and alternate between maximizing posterior distributions of Θ_u and using the above results to update prior parameters Λ_u and β_u until convergence criterion is satisfied.

The framework of empirical Bayes guarantees the precision of these learning variables. In this case, we need to explain why we do not use *fully Bayesian* treatment. As we discussed above, all variables are assumed to be Gaussian, which renders a closed form when making inferences and predictions. However, if we adopt the framework of fully Bayesian treatment, both inferences and predictions are analytically intractable. Therefore, we need to resort to approximate inference like variational methods [5] or MCMC-based methods [8]. Variational methods typically scale well to large applications and may produce inaccurate results for our estimation problem, while MCMC-based methods are too time-consuming for training our model. Considering this, we employ the framework of empirical Bayes, which can save plenty of time for inferences and predictions.

5 Experimental Results

In this section, we comprehensively evaluate our procrastination discovery approach based on several real-world data sets. First, we evaluate the effectiveness of our models through predicting the hold time of borrowed books. Then, we empirically verify our learned procrastination value from psychological fields.

5.1 Prediction Performance

Experimental Setup. In our experiments, we empirically extracted 812,506 records from 10,035 students, who borrowed more than 30 books throughout the four years in college, as evaluation data set. Furthermore, we randomly selected 144,590 records as test set and the remaining records were used for training models. To the best of our knowledge, there is no existing work for the hold time of borrowed books prediction. Therefore, we exploit several intuitive but state-of-the-art baselines for evaluation. First, we propose to use the the average hold time of previous borrowed books of the given student for prediction (i.e., *Average*), which is based on the intuition of habitual momentum. Second, from the preference factorization perspective, we use the Probabilistic Matrix Factorization (i.e., *PMF*) [7] with 30 latent factors for predicting the hold time. Third, we choose some supervised machine learning techniques for prediction. The selected features for these methods are listed in Table 2.

In our experiments, we exploited Weka to conduct the above baselines, which is an open source software under the GNU General Public License[1]. The parameters of these machine learning methods and the value of λ in our naive RP model are set according to 10-fold cross validation. Besides, we empirically set K to be 3 in dynamic RP model. The time of *kernel points* were set to September 15th, November 15th and January 15th for the first semester, while March 15th, May 15th and July 15 for the second semester, since they correspond to the start, middle and end of each semester, respectively.

[1] http://www.cs.waikato.ac.nz/ml/weka/.

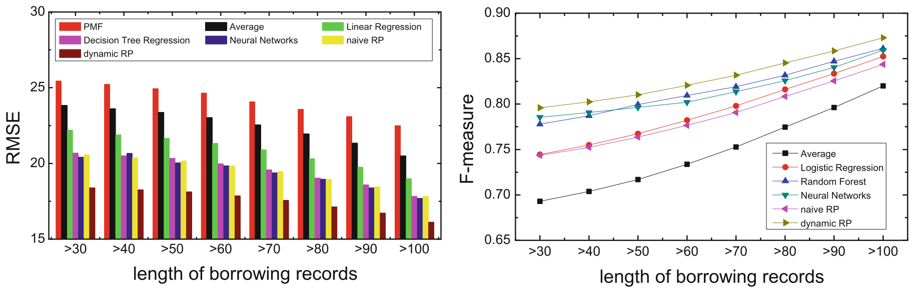
Table 2. Selected prediction features

For student features:
• Borrowing number previously
• Minimum hold days of borrowed books previously
• Maximum hold days of borrowed books previously
• Average hold days of borrowed books previously
• Hold days of borrowed book last time
• Number of times coming to library previously
For book features:
• Number of times been borrowed
• Minimum hold days been borrowed
• Maximum hold days been borrowed
• Average hold days been borrowed
For interaction features:
• Borrowing semester
• Borrowing month

Evaluation Metrics. For predicting the hold time of borrowed books, we can treat it as a regression problem or classification problem. The regression performance can be evaluated by the Root Mean Squared Error (RMSE), which is defined as $RMSE = \sqrt{\frac{\sum_{(u,i)\in C}(h_{u,i}(t) - \hat{h_{u,i}}(t))^2}{|C|}}$, where $\hat{h_{u,i}}(t)$ is the predicted hold time of the real value $h_{u,i}(t)$, and C denotes the test set. As for classification performance, we can separate the records into two classes for prediction, where one represents those returned back within a month (i.e., 30 days) and the other represents those longer than a month. We regard the books returned back within 30 days as positive class and leverage the classic evaluation metric *F-measure* for evaluation, which is harmonic mean of metric *Precision* and *Recall.*

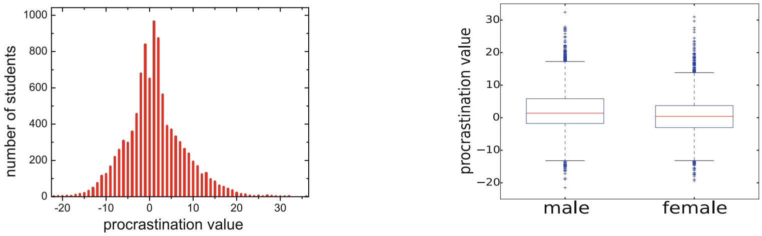
Performance and Discussion. Fig. 3 shows the RMSE performance and F-measure performance with respect to students who borrowed different number of books totally. From the two figures, we can observe that our dynamic RP model has the best prediction performance. As for the naive RP model, it has comparative performance with the machine learning methods. Particularly, more attention should be paid to the performance of *PMF*, which has the highest value of RMSE, even higher than the *Average* method. This indicates that the relationship between students and books based on the latent preferences are not important for books hold time.

Based on the above experimental results, we can obtain the following conclusions. First, our two proposed models are effective in predicting the hold time of books borrowed by college students, especially the dynamic RP model. This guarantees the validity of our procrastination assumption. Although there still



(a) (b)

Fig. 3. Performance with respect to students who borrowed different number of books. (a) Regression performance. (b) Classification performance.



(a) (b)

Fig. 4. Distribution of the learned procrastination value. (a): The number of students with respect to the procrastination value. (b): The distribution of procrastination value with to male and female students.

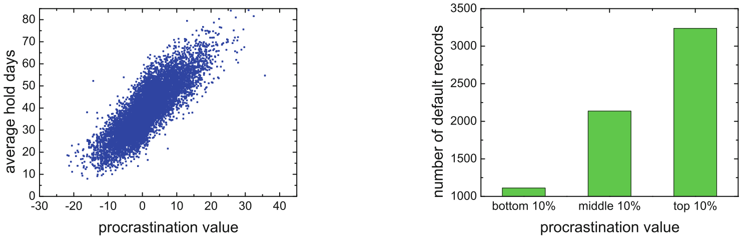
exists some predictive bias, two reasons can explain this result. One is that the events happened in a short period of time for each student are hardly captured, which can also affect the hold time of borrowed books more or less. The other is that there are still some errors in the estimation of *required reading time*, which is also difficult to seek the optimal estimation. Second, some machine learning methods such as *Decision Tree Regression* and *Neural Networks* have nearly same regression performance with the naive RP model. This implies that these machine learning methods are very likely to discover the knowledge about the *required reading time* and procrastination factors from data, while it is difficult for them to capture more dynamic factors. Moreover, the lack of the ability to discover the procrastination of students is the crucial limitation, which is the superiority of our models.

5.2 Procrastination Verification

In this subsection, we validate our learned procrastination value of students.

Distribution Evidences. Figure 4 shows the distribution of the procrastination value of all 10,035 students in our data set. Specifically, Fig. 4(a) illustrates the number of students with respect to the learned procrastination value. The greater the procrastination value is, the heavier extent of procrastination the student has. As for students who have negative procrastination value, we explain them the opposite of procrastination, the sense of urgency. From this figure, we can observe the normality of our learned procrastination value. Figure 4(b) reveals the distribution of procrastination value with respect to male and female students. As reported by Van et al. [17], it is slightly more likely that men procrastinate more than women. From this figure, we can notice that male students have relatively higher procrastination value than female students, and the mean value of male and female students are 2.05 and 0.57 respectively. This consistent with psychological result evidences the truth our learned procrastination value.

Library Actions Evidences. To illustrate how procrastination makes impact on the library actions. Figure 5(a) provides the relationship between the procrastination value and the average hold days of all borrowed books over four years in college for corresponding students. In this figure, there exists a strong trend that the average hold days are increasing with the increase of procrastination value. This indicates that students with higher procrastination value are more inclined to hold borrowed books for a longer time. Besides, it also has correlation between the procrastination behavior and default behavior in college library. Here, we have 32,648 default records of penalty in our library data set. We selected the bottom 10 %, middle 10 % and top 10 % students according to the magnitude of procrastination value. Figure 5(b) presents the number of default records with these students. In this figure, apparently, the top 10 % students with the highest procrastination value have much more default records, whereas the bottom 10 % students have much less. It is nature that those top 10 % students procrastinate to return their borrowed books, then gradually evolve to forget it and exceed the due return time. According to the above statistics, the rationality of our learned procrastination value can be verified. Besides, as an application of library service, students with high procrastination level will be reminded to return the borrowed books in case these books are in urgent need by other students.



(a) (b)

Fig. 5. The statistics for the procrastination value with library actions. (a): The average hold days of all borrowed books over four years in college for students with corresponding procrastination value. (b): The number of default records in library with the bottom 10 %, middle 10 % and top 10 % procrastination value.

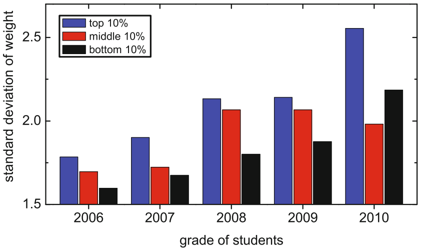


Fig. 6. The average standard deviation of weight over four years in college with the bottom 10 %, middle 10 % and top 10 % procrastination value, grouped by entering college year.

Association Evidences. To further verify our learned procrastination value, we take the association results between procrastination value and other characteristics of students with psychological fields conclusions for comparison. As Culnan et al. [3] revealed that procrastination may influence college freshmen weight change, we first test our learned procrastination value on students' body weight data. Figure 6 shows the average standard deviation of students weight over four years in college with the bottom 10 %, middle 10 % and top 10 % procrastination value. Here, we group students according to their entering college year. From this figure, we can find that the weight of students with highest procrastination value fluctuates more greatly over four years in college than that of students with lowest procrastination value. This statistical result is consistent with Culnan's conclusion, which evidences the reliability of our learned procrastination value. As for academic performance, Steel et al. [12] noted that procrastination may not have contributed significantly to poorer grades and students who completed all of the practice exercises tended to perform well on the final exam no matter how much they delayed. Therefore, we test on the scholarship data. There are 1,267 students winning a scholarship, and the number of students having procrastination value in the bottom 10 %, middle 10 % and top 10 % account for 58, 70 and 76 respectively. We find that the gap between top 10 % and bottom 10 % or middle 10 % is not obvious, it seems that the number of students with higher procrastination value is even more. This consistent statistics also gives more evidence to the validity of our learned procrastination value.

6 Related Work

Generally, the related work of this paper can be grouped into two categories, i.e., procrastination in psychological research and behaviors analysis of students.

Procrastination in Psychological Research. Procrastination has been studied for a long time in the field of psychology. One of research directions is the study of impact brought by procrastination behavior. Tice et al. [16] found that some negative associations are linked to procrastination, such as depression, anxiety, irrational behaviour and low self-esteem. Another research direction focuses on the exploration of the possible causes of procrastination. For example, Steel et al. [11] revealed that task aversiveness, task delay, self-efficacy, and impulsiveness, as well as conscientiousness and its facets of self-control, distractibility, organization, and achievement motivation are strong and consistent predictors of procrastination. Study also involves in result of the remedy of procrastination behavior. For example, Ariely et al. [2] specially studied people who strategically try to curb procrastination by using costly self-imposed deadlines. Their empirical evidence showed that self-imposed deadlines are not always as effective as some external deadlines in boosting task performance.

Behaviors Analysis of Students. The behaviors analysis of students is attracting more researchers these years due to the increasing available data. Recently, Guan et al. [4] developed a learning framework Dis-HARD for identifying students who are qualified to obtain the financial funding support in college by investigating student's complex behaviors within campus. Agrawal et al. [1] proposed solutions for grouping students who exhibit different ability level into sections so that the overall gain for students is maximized. To examine students' learning process and improve their study performance, researches also covered in education, emerging educational data mining (EDM) [9].

Above all, however, no existing work has explored the procrastination from college library data, which comes out our procrastination exploration work.

7 Concluding Remarks

In this paper, we introduced a data-driven study from a behavioral perspective to explore the procrastination of college students. To this end, we proposed an unsupervised approach to quantitatively estimate the procrastination level of students through the analysis of their borrowing records in college library. Specifically, we first propose a naive Reading-Procrastination (naive RP) model, which takes consideration of the behavioral similarity between students for procrastination discovery. Furthermore, to improve the discovery performance, we develop a dynamic Reading-Procrastination (dynamic RP) model by integrating more comprehensive characteristics of student behaviors, such as semester-awareness and month-regularity. A unique characteristics of the dynamic RP model is it can depict the procrastination behavior in a probabilistic and empirical Bayesian perspective. Finally, we conducted extensive experiments on several real-world data sets collected from a Chinese college. The experimental results clearly demonstrated the effectiveness of our approach, and verified several key findings from psychological fields.

Acknowledgments. This research was partially supported by grants from the National Science Foundation for Distinguished Young Scholars of China (Grant No. 61325010), the Science and Technology Program for Public Wellbeing (Grant No. 2013GS340302). Qi Liu gratefully acknowledges the support of the Youth Innovation Promotion Association of CAS and acknowledges the support of the CCF-Intel Young Faculty Researcher Program (YFRP).

References

1. Agrawal, R., Golshan, B., Terzi, E.: Grouping students in educational settings. In: SIGKDD 2014, pp. 1017–1026. ACM (2014)
2. Ariely, D., Wertenbroch, K.: Procrastination, deadlines, and performance: self-control by precommitment. Psychol. Sci. **13**(3), 219–224 (2002)
3. Culnan, E., Kloss, J.D., Grandner, M.: A prospective study of weight gain associated with chronotype among college freshmen. Chronobiol. Int. **30**(5), 682–690 (2013)
4. Guan, C., Lu, X., Li, X., Chen, E., Zhou, W., Xiong, H.: Discovery of college students in financial hardship. In: ICDM. IEEE (2015)
5. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. Mach. Learn. **37**(2), 183–233 (1999)
6. Klassen, R.M., Krawchuk, L.L., Rajani, S.: Academic procrastination of undergraduates: low self-efficacy to self-regulate predicts higher levels of procrastination. Contemp. Educ. Psychol. **33**(4), 915–931 (2008)
7. Mnih, A., Salakhutdinov, R.: Probabilistic matrix factorization. In: Advances in neural information processing systems, pp. 1257–1264 (2007)
8. Neal, R.M.: Probabilistic inference using markov chain monte carlo methods. Technical report CRG-TR-93-1 (1993)
9. Romero, C., Ventura, S.: Data mining in education. Wiley Interdisc. Rev.: Data Min. Knowl. Discov. (DMKD) **3**(1), 12–27 (2013)
10. Solomon, L.J., Rothblum, E.D.: Academic procrastination: frequency and cognitive-behavioral correlates. J. Couns. Psychol. **31**(4), 503 (1984)
11. Steel, P.: The nature of procrastination: a meta-analytic and theoretical review of quintessential self-regulatory failure. Psychol. Bull. **133**(1), 65 (2007)
12. Steel, P.: The Procrastination Equation: How to Stop Putting Things Off and Start Getting Stuff Done. Random House Canada, Toronto (2010)
13. Steel, P., Brothen, T., Wambach, C.: Procrastination and personality, performance, and mood. Personality Individ. Differ. **30**(1), 95–106 (2001)
14. Sub, A., Prabha, C.: Academic performance in relation to perfectionism, test procrastination and test anxiety of high school children. Psychological Studies (2003)
15. Tan, P.N., Steinbach, M., Kumar, V., et al.: Introduction to Data Mining, vol. 1. Pearson Addison Wesley, Boston (2006)
16. Tice, D.M., Baumeister, R.F.: Longitudinal study of procrastination, performance, stress, and health: the costs and benefits of dawdling. Psychol. Sci. **8**, 454–458 (1997)
17. Van Eerde, W.: A meta-analytically derived nomological network of procrastination. Personality Individ. Differ. **35**(6), 1401–1418 (2003)
18. Wikipedia: Procrastination – wikipedia, the free encyclopedia (2015). http://en.wikipedia.org/wiki/Procrastination. Accessed 1 May 2015