

How Surfing Habits Affect Academic Performance: An Experimental Study

Xing Xu^{1,2(✉)}, Jianzhong Wang¹, and Haoran Wang³

¹ School of Humanities and Social Sciences, Beihang University, Beijing, China
{xuxing66,wjz}@buaa.edu.cn

² Network Information Center, Beihang University, Beijing, China

³ School of Software, Beihang University, Beijing, China
18811399416@163.com

Abstract. The issues regarding relationships between surfing habits and academic performance of university students have attracted much attention of all kinds of research communities. The approaches of statistical analysis and data mining with imperfect datasets are used in previous works. In this paper, an experimental study about the relationships between surfing habits and academic performance is conducted. Particularly, we observe a surprising results on extensive datasets, which contains the information of students, e.g. basic profiles, Internet using logs and course scores. First, several statistical methods are used to find the associations between students' surfing habits and academic performance. Then, a learning algorithm is devised to cluster the students according to their different surfing habits. Furthermore, we develop a BP neural network to predict the rate of failing a test of a student based on his/her basic information, surfing habits and the clustering information. According to the aforementioned approaches, we find an interesting result that the academic performance of students in universities is quite possibly enhanced if the network connections are always terminated at 24:00 p.m.

Keywords: Internet surfing habits · Course score · Exam-failure rate · Cluster analysis · Neural network analysis

1 Introduction

According to the Statistics of Chinese Ministry of Education in 2016 [1], there are 2,852 universities and colleges in China, and the number of university students reaches 3.7 billion, which is the most around the world. With the development the Internet, it is convenient for university students to use it in almost everywhere in the campus through PCs, Laptops, PADs, Smart phones and so on. The Internet affects every aspect of the life of students including learning, entertainment, social activity, etc. Students are encouraged to communicate with teachers, instructors and classmates through the Internet and search information on it.

Some researches indicate that academic performance can be promoted with the help of the Internet. Rashid and Asgharb suggests that the use of the Internet has a direct and positive influence on students' engagement and self-directed learning [9]. Dr. Suhail and Bargees's investigation finds that a great number of students demonstrate positive effects of Internet using on educational aspect [13]. In the U.S., 68 % of the parents, 69 % of the students, and 69 % of the teachers say that they have authenticated the improvement of students' grades through the use of the Internet on an online survey reports [2]. In addition, the Internet contributes significantly to the academic performance of the students in university of Botswana, Gaborone, possibly because the students can learn at anytime and anywhere with the Internet [14]. The broad use of the Internet in universities can make students have better perceptions of learning [4].

However, there are also some researches indicating that the Internet using has negative effects. General Internet surfing habits of students show that some people may experience psychological problems such as social isolation, depression, loneliness, and time mismanagement because of their Internet using [7]. Bad academic performance has been confirmed to be correlated highly with heavier recreational Internet using. Loneliness, staying up late, tiredness, and missing classes are also correlative to self-reports of Internet-caused impairment [8]. Heavy Internet usage may have bad relationships with teachers, bad academic grades and low learning satisfaction than general users. And they are apt to be depressed, physically ill, lonely, and introverted [3]. Especially, with the increasing popularity of smart phone, more people start relying on it. The data from the Pew Research Center in 2015 shows that 46 % of smart phone owners in the U.S. say that their smart phone is something "they could not live without". Smart phones are used all day for quite a lot of reasons, including communication, learning, entertainment, information seeking, social networking and gaming [17]. In particular, recent research shows that online social communications based on smart phones significantly affect offline behavior of users [10–12, 18–20]. University students who are at a high risk of smart phone addiction may be less likely to have good GPAs [5]. In addition, excessive use of the social network systems results in negative effects. For example, multitasking with Facebook has been found to be significantly negatively predictive of GPAs for freshmen and sophomores [6].

The correlation between students' Internet using and their academic performance is a complex issue. Average daily connection time, as a general used metric to determine Internet addiction, may not be accurate because students may study using the Internet for a long time other than being addicted to online entertainment. In fact, Internet addiction of students in universities is reflected not only by Internet using time but also by other features such as online and offline time, volume of Internet traffic and other potential factors. The relationships between these factors and the performance of students are very complex. In addition, students' background features, including gender, major, hometown, etc., differ from one another, which may lead to different online behaviors. All these features should be taken into consideration when evaluating the impact of

Internet using habits on academic performance of university students. We make the following contributions.

- We reconfirm that no obvious relationship between the Internet using habits and the academic performance can be found using linear regression.
- We construct a model to predict whether a student will fail a test using the Internet using and other features.
- Based on the prediction model, we make a suggestion that the network connections should be terminated at 0:00 to improve students' academic performance.

The rest of the paper is organized as follows. In Sect. 2, we introduce the datasets and the methods used in this paper. In Sect. 3, we give the analysis results. We conclude the paper in Sect. 4.

2 Preliminaries

2.1 Datasets

In order to get accurate results, the data employed in this study are obtained from several source providers, e.g., the Internet using data are obtained from ITS (IT Service Department) of the university, the student academic performance data are obtained from the Academic Affairs office, and the students' basic data (background features) are obtained from the Admission department. There are four grades of students, ranging from freshman to senior college students. Students' basic dataset has 23,175 records, Internet using dataset has 321,651,912 items, and academic performance dataset has 510,261 courses information records. Each record of the students' basic dataset contains the following attributes: ID number, gender, grade, school, major and source province. Each record of the Internet using dataset contains the following attributes: ID number, online time, offline time, Internet traffic volume (bit) and connection time. Each record of the academic performance dataset contained the following attributes: ID number, course number, course grade, year and semester.

2.2 Data Preparation and Statistics

A record of in the students' basic dataset is represented as $(d_1, p_1, b_1, m_1, e_1)$, where d_1 is the key of the student's basic information table, $p_1 \in P$ is the province from which the student comes, b_1 is the birthday, m_1 is the major and $e_1 \in E$ is the nationality. P and E are the collections of provinces and nationalities in China, respectively.

A record in the Internet using habit dataset is represented as (d_2, l_2, f_2, b_2) , where d_2 is the key of the table, l_2 and f_2 is the online time and the offline time respectively and b_2 is the Internet traffic volume. Two important features included to reflect the Internet using habit of the students are the online time and the Internet traffic volume.

A record in the dataset of student grade is represented as (d_3, c_3, g_3) , where d_3 is the unique ID of the student in this table, c_3 is the course ID, and g_3 is the examine score of the corresponding course.

2.2.1 Internet Traffic Volume per Week of Students

We use the Internet using habit data for statistics, and get the weekly Internet traffic volume of the university students. The average traffic volume per week of a student is about 9 Gb. The average weekly traffic volume is divided into seven categories: (1) less than 5 Gb, (2) 5–10 Gb (excluding 10 Gb), (3) 10–15 Gb (excluding 15 Gb), (4) 15–20 Gb (excluding 20 Gb), (5) 20–25 Gb (excluding 25 Gb), (6) 25–30 Gb (excluding 30 Gb), (7) no less than 30 Gb. 36.5% of the students use 10–15 Gb Internet traffic volume weekly, taking the majority. 24.3% of the students use 15–20 Gb Internet traffic volume, which is the second most, and 5–10 Gb followed as the third. Only 2.2% of the students use no less than 30 Gb per week. More than 60% of the students' weekly Internet traffic volume is between 10 Gb and 20 Gb. Figure 1 shows the distribution of Internet traffic volume per week.

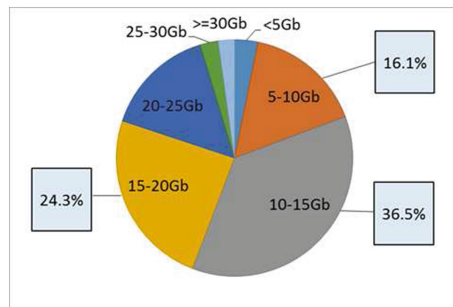


Fig. 1. The distribution of Internet traffic volume per week

2.2.2 Internet Connection Time of Students

We use average Internet connection time per week of a student as a factor. The average value of all students is 32 hours. Figure 2 shows the distribution of average Internet connection time per week. similarly, we divide the time into seven intervals: (1) less than 10 h, (2) 10–20 h (excluding 20 h), (3) 20–30 h (excluding 30 h), (4) 30–40 h (excluding 40 h), (5) 40–50 h (excluding 50 h), (6) 50–60 h (excluding 60 h), (7) no less than 60 h. From the figure it can be observed that most students keep online for 20–50 h per week. The top-3 intervals with the most students are 30–40 h, 20–30 h and 40–50 h, taking the percentage of 37.3%, 25.6% and 19.1% respectively. More than 1/3 of the students keep online for 30–40 hours per week. It seems that the students are most likely to connect Internet for 20 to 50 h every week, which accounted for about 82% in the general situation. The distribution of Internet using time in one-day (24 h) is depicted as Fig. 3. 11:00 a.m., 20:00 p.m. and 24:00 p.m. takes the leading places, and the accumulative quantities of the Internet using time at 11:00, 12:00, 13:00, 20:00, 21:00, 23:00, 24:00 are more than 50% of the total. After 24:00 p.m., about

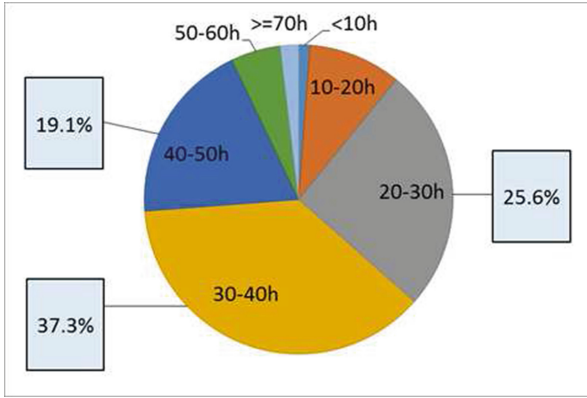


Fig. 2. The distribution of average weekly Internet connection time

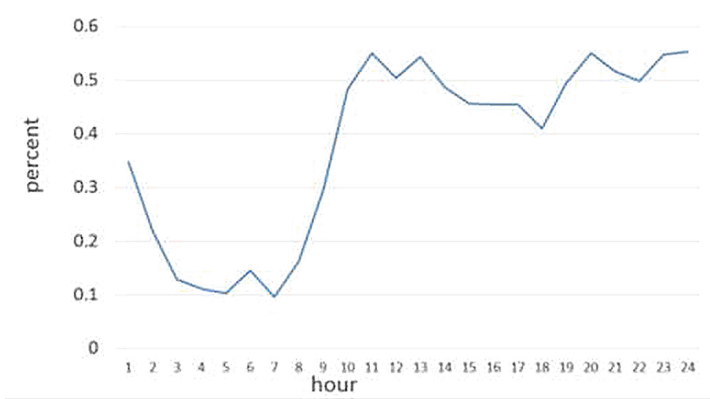


Fig. 3. The distribution of average weekly Internet connection time

20 % of the students disconnect the Internet for sleeping, but there are still more than 20 % of the students keeping online. From 3:00 to 7:00 a.m., the Internet connection quantities reach the lowest level.

2.2.3 Academic Performance of the Students

Figure 4 shows the distribution of the academic performance. According to the scores the students obtain, grades are divided into five parts: (1) less than 60, (2) 60–70 (excluding 70), (3) 70–80 (excluding 80), (4) 80–90 (excluding 90), (5) no less than 90. The maximum score on all the tests is 100. The scores of more than 1/3 of all students range from 80 to 90. The rate of failing an exam is 13.7%.

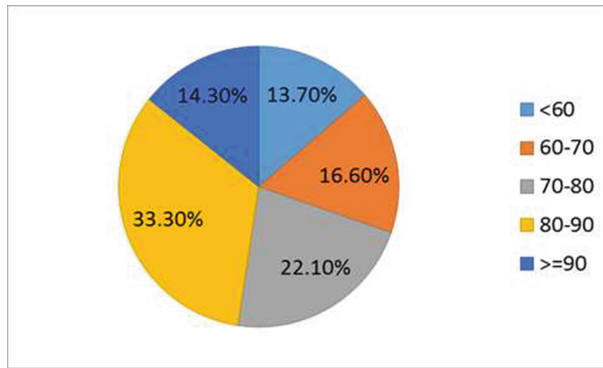


Fig. 4. Academic performance of students

2.2.4 Relationship Between Connection Time and Academic Performance

Will academic performance be determined by Internet using? Do the 1.9% of the students who connect the Internet more than 60 h per week always fail the exams? The distribution of the academic performance based on the connection time is showed in Fig. 5. The outliers should be neglected, and it seems that there is no obvious characteristic of their relationship. The students with different connection time have similar distribution of scores. In order to do further analysis, the linear regression is used to analyze the relationship between the Internet connection time and the academic performance. We get the correlation coefficient of 0.016321. This result suggests that they no significant correlationship.

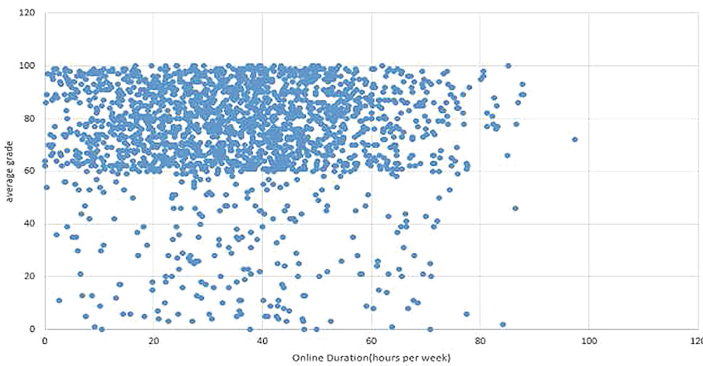


Fig. 5. The distribution of academic result based on Internet connection time

2.3 The System Framework

However, due to the data complexity, the relationship between the Internet using habit and the academic performance cannot be analyzed simply by calculating the correlation between the exam performance and the Internet connection time. More features should be taken into consideration. Therefore, multivariate fitting is introduced to get a higher degree of fitting relationship. In order to get more accurate relationship, the techniques of clustering and neural network are used. The workflow of the framework is showed in Fig. 6. Three types of datasets including the Internet using data, the academic performance data and the students' basic data, are used to extract the features of students for clustering. Then we use neural network to fit with the clustering results. The recommended disconnection time is calculated through continuously automatic fitting.

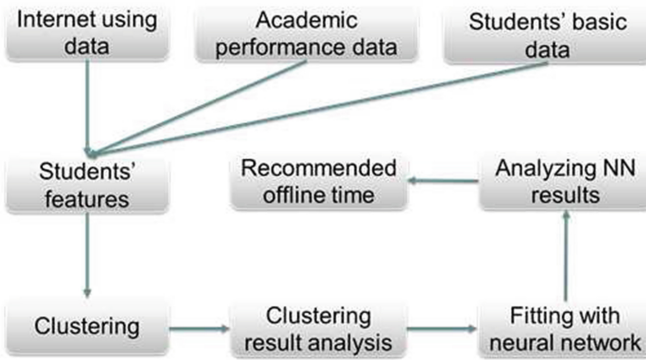


Fig. 6. Process of the framework

2.4 The Data Models

The three datasets which are the Internet using data, the academic performance data and the student basic data, are employed to extract the following features: Internet using habit, source province, gender, major, connection time per week and Internet traffic volume per week.

Internet using habit is defined as the connection time series of one-day. 24 elements $(t_0, t_1, t_2, t_3, t_4, \dots, t_{23})$ are contained, where, $t_i (i = 0, 1, 2, \dots, 23)$ represents the number of days, (d_i) that a student is online during a certain hour of day (i.e. the i th hour), T is the total days of sampled days(a year, 365 days):

$$t_i = d_i/T \tag{1}$$

Every 4 continuous hours in a day is used to be a period as the factors to run clustering algorithms. The items are expressed as $(f_1, f_2, f_3, f_4, f_5, f_6)$, and

$$f_j = \frac{\sum_{l=4(j-1)}^{4j-1} t_l}{4} \tag{2}$$

A student's performance in courses is represented by a boolean value where 0 represents passing the course and 1 represents failing. P is the score of course. C is the collection of all the courses studied in the sample semesters

$$P = \begin{cases} 0, \forall p \in C, p \geq 60 \\ 1, \exists p \in C, p < 60 \end{cases} \quad (3)$$

2.5 Cleansing Sensitive Data

To get accurate result, the noise and uncertain data should be eliminated firstly. For example, the records of very short connection time or very low Internet traffic volume should be removed. Since these sensitive data are infrequent, we use the uncertain frequent itemset mining techniques [15, 16, 21] to discover infrequent records. Although some data reach the selection standard of the Internet connection time threshold and the Internet traffic volume threshold, their periods (number of days) are relatively low, possibly less than one month during the two-year period. These discontinuous data should not be used to represent long-term behavior, and is removed as well.

2.6 Clustering Analysis

We use DB-SCAN as the clustering algorithm. A 6-dimensional vector is defined as the input, the smallest class cluster is limited within 1,500, and the scanning radius is 0.15623. Finally, 15,062 results data and 11,023 noise data are obtained during a year. The result containing six clusters is depicted in Fig. 7, which shows the line chart of the Internet habit of the six clusters. The before-dawn online period (from 1 am to 6 am) of clusters 4, 5 and 6 are nearly 0. For cluster 1, 2, and 3, the data may refer to the Internet online records of several students. Comparing to other periods, the Internet connections of this period in these clusters are also low, although, the percentage of the average connection is approximately 5%. 20:00 p.m. is the peak of the average online time. At this time point, the online proportions are the highest among most clustered groups. For each cluster, the Internet using habits of morning and afternoon are different. Some fluctuates largely, while the others are relatively steady.

2.7 Classification Analysis

These three datasets in that year are used for classification analysis. There are 15,062 records of the students' basic information and the Internet using habit data, among which, 12,000 records of data are used as training set, and 2,500 used as testing set for accuracy verification with cross validation. Classification analysis is carried out using BP Neural Network. BP Neural Network is composed of input layer, hidden layers and output layer. The structure of the BP Neural Network is showed in Fig. 8. In this study, 7 features of each student are used in a BP neural network training, and the output layer node represents the exam failure. The used 7 features are: the student's Internet habit, source province,

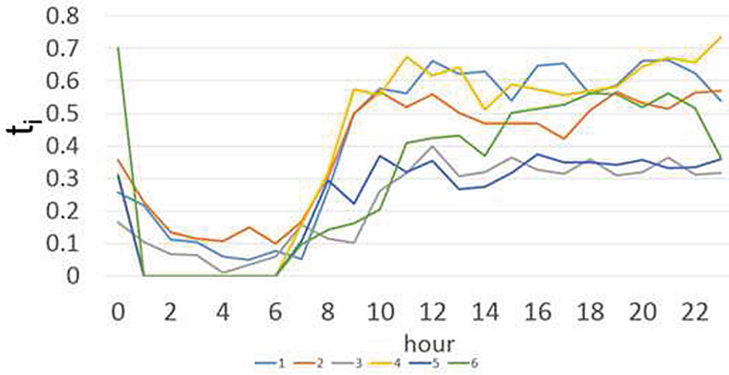


Fig. 7. Internet habit of 6 clusters

gender, major, weekly Internet connection time, weekly Internet traffic volume, and the cluster ID which the student belongs to. The offline in evening may be the best time for improving exam score, so we make the hypothesis that adjusting the offline time in evening will help reducing the exam failure rate. The student’s Internet habit is represent with nighttime Internet using habits ($t_{21}, t_{22}, t_{23}, t_0, t_1, t_2, t_3, t_4, t_5$). Thus, The input layer is divided into 15 nodes: province, major, gender, age, Internet traffic volume, Internet connection time, and Internet using habit which comprises ($t_{21}, t_{22}, t_{23}, t_0, t_1, t_2, t_3, t_4, t_5$). The output layer is exam failure, where, 1 represents that the student has failed at least one test of the course, and 0 indicates that the student has no exam failure.

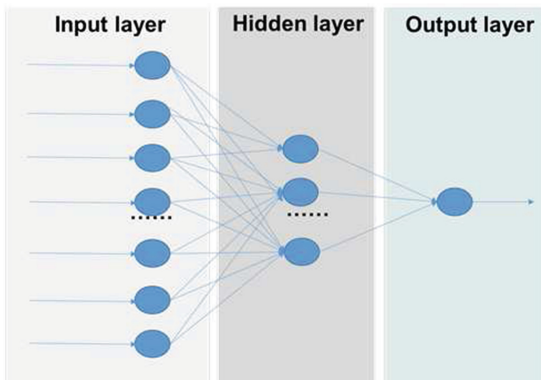


Fig. 8. The structure of the BP Neural Network

With 15 input layer nodes, and 1 output layer node, according to the empirical formula, the quantity of the hidden layer nodes ranges from 5 to 14. Making the largest study iteration number as 5,000, and under the accuracy condition

of 0.1, the results indicates that with 11 hidden layer nodes, the optimal effect reaches 91.625%. So, accuracy of the hidden layer increases when the nodes added from 5 to 11, while decreases with nodes from 11 to 14.

After the establishment of neural network modeling, the model is used to predict the effect of the disconnection time on students' exam failure rate. Five groups of students are selected randomly from cluster 4, 5, 6. Each group has 500 students, and their actual failure rates are as shown in Table 1.

Table 1. Failure rate of each group

Group number	Failure rate
1	0.132
2	0.126
3	0.112
4	0.122
5	0.126

3 The Experimental Results

3.1 Disconnection Time v.s. Exam Failure

By changing $t_{21}, t_{22}, t_{23}, t_0, t_1$ to 0 respectively, the simulation data is obtained and shown in Fig. 9. The failure rates of the five groups is declining smoothly till 24:00 p.m., and reach the lowest at 24:00 p.m. It indicates that there may be some students use Internet for course learning before 24:00 p.m, and hence 24:00 p.m. should be the time to disconnect the Internet. If the students keep online till 1:00 a.m., the failure rates increase sharply. So taking the 24:00 p.m. as the disconnection time would improve students' academic performance in this university.

3.2 The Influence of Other Factors

In addition to online time, there are other elements being used for neural network modeling to predict the students' academic performance. Source provinces and academic majors have some influence on academic performance. Hebei province, Hunan province and Hubei province where students come from are the top 3 provinces of reduction of exam failure rate in response to adjusted offline time (12:00 p.m.). Similarly, students who major in mechanics, transportation, and Liberal arts show a similar trend in their academic performance in response to the adjusted time. However, the difference between students of Hans and other nationalities is not obvious. That is to say the nationality of students has nothing to do with the Internet connection limitations which would lead to decline of the exam-failure rate.

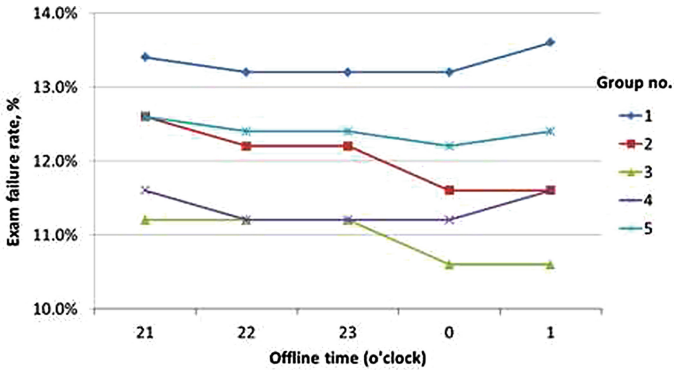


Fig. 9. The failure rate of the five group

3.3 Incorrect Prediction Explanation

About 9% of the incorrect predictions mainly come from the insufficient data acquisition and condition prediction. Even though neural network has high multivariate fitting effect, the analysis of students' academic performance has high complexity, and factors with the students' attributes of this study are not enough.

4 Conclusion

With the popularity of campus network service, it is meaningful to research the relationship between the Internet using habits and the academic performance of university students. A comprehensive dataset of the past five years from an university in China are analyzed in this paper. Firstly, statistical methods are used, and we reconfirm that no obvious relationship between the Internet using habits and the academic performance can be found with linear regression. Then, data mining methods are applied. The clustering algorithm DB-SCAN is used to classify the online users, and a BP neural network is trained for deep mining. Lastly, the best moment of 24:00 p.m. is gotten by simulating the disconnection time with the trained BP neural network model. Therefore, in order to improve students' academic performance, we suggest that the Internet should be disconnected at 24:00 p.m.

References

1. <http://www.moe.gov.cn>
2. <http://www.att.com/press/0998/980902.csa.html>
3. Chen, Y., Peng, S.: University students' internet use and its relationships with academic performance, interpersonal relationships, psychosocial adjustment, and self-evaluation. *CyberPsychol. Behav.* **11**(4), 467–469 (2008)

4. Cheung, W., Huang, W.: Proposing a framework to assess internet usage in university education: an empirical investigation from a student's perspective. *Br. J. Educ. Technol.* **36**(2), 237–253 (2005)
5. Hawi, N., Samaha, M.: To excel or not to excel: Strong evidence on the adverse effect of smartphone addiction on academic performance. *Comput. Educ.* **98**, 81–89 (2016)
6. Junco, R.: Student class standing, facebook use, and academic performance. *J. Appl. Dev. Psychol.* **36**, 18–29 (2015)
7. Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukophadhyay, T., Scherlis, W.: Internet paradox: A social technology that reduces social involvement and psychological well-being? *Am. Psychol.* **53**(9), 1017 (1998)
8. Kubey, R., Lavin, M., Barrows, J.: Internet use and collegiate academic performance decrements: Early findings. *J. Commun.* **51**(2), 366–382 (2001)
9. Rashid, T., Asghar, H.: Technology use, self-directed learning, student engagement and academic performance: Examining the interrelations. *Comput. Hum. Behav.* **63**, 604–612 (2016)
10. She, J., Tong, Y., Chen, L.: Utility-aware social event-participant planning. In: *SIGMOD 2015*, pp. 1629–1643 (2015)
11. She, J., Tong, Y., Chen, L., Cao, C.C.: Conflict-aware event-participant arrangement. In: *ICDE 2015*, pp. 735–746 (2015)
12. She, J., Tong, Y., Chen, L., Cao, C.C.: Conflict-aware event-participant arrangement and its variant for online setting. *IEEE Trans. Knowl. Data Eng.* **28**(9), 2281–2295 (2016)
13. Suhail, K., Bargees, Z.: Effects of excessive internet use on undergraduate students in pakistan. *CyberPsychol. Behav.* **9**(3), 297–307 (2006)
14. Tella, A.: Undergraduates uses of the internet: implications on academic performance. *J. Educ. Media Libr. Sci.* **45**(2), 161–185 (2007)
15. Tong, Y., Chen, L., Cheng, Y., Yu, P.S.: Mining frequent itemsets over uncertain databases. *Proc. VLDB Endow.* **5**(11), 1650–1661 (2012)
16. Tong, Y., Chen, L., Ding, B.: Discovering threshold-based frequent closed itemsets over probabilistic data. In: *ICDE 2012*, pp. 270–281 (2012)
17. Tong, Y., She, J., Chen, L.: Towards better understanding of app functions. *J. Comput. Sci. Technol.* **30**(5), 1130–1140 (2015)
18. Tong, Y., She, J., Ding, B., Chen, L., Wo, T., Xu, K.: Online minimum matching in real-time spatial data: experiments and analysis. *Proc. VLDB Endow.* **9**(12), 1053–1064 (2016)
19. Tong, Y., She, J., Ding, B., Wang, L., Chen, L.: Online mobile micro-task allocation in spatial crowdsourcing. In: *ICDE 2016*, pp. 49–60 (2016)
20. Tong, Y., She, J., Meng, R.: Bottleneck-aware arrangement over event-based social networks: the max-min approach. *World Wide Web J.* **19**(6), 1151–1177 (2016)
21. Tong, Y., Zhang, X., Chen, L.: Tracking frequent items over distributed probabilistic data. *World Wide Web J.* **19**(4), 579–604 (2016)