Mining gender affinity to social networking

Rozita Jamili Oskouei*
Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology, Allahabad, UP, 21004, India
E-mail: rosejamili2009@gmail.com
E-mail: rozita2010r@gmail.com
*Corresponding author

Phani Sajja
Expert Software Consultants Ltd., B-77-78, 2nd Floor, Sindhuja Building, Sector – 6, NOIDA, UP, 201 301, India
E-mail: sajja.phani@gmail.com

Abstract: Social networking websites are currently used by many people, especially young ones who are in the age group of 17–25. Since these age groups are mostly students and are either college going or in the process of making a career, it is difficult for the parents to keep a close watch over their activities. Even the academic coordinators and advisors have not developed a pattern through which these age groups can be guided and trained. Nevertheless, it is difficult to pattern the personal and social behaviour, academic activities and performances. There is no exhaustive study or research work done to address this issue in different countries.

In this research, a comprehensive analysis of students' internet usage behaviours by focusing on social network (SN) website usages is targeted. It is an attempt to pattern the characteristics by demonstrating and elaborating the various aspects of Internet usages. Relationship between users' usage patterns and their personal and social behaviours along with their academic performances are studied, finally a prediction method for predicting students' future academic performance by analysing their internet and SN websites usage patterns is proposed.

Keywords: behaviour mining; internet usage pattern; social network; web usage mining; WUM; website classification.

Reference to this paper should be made as follows: Oskouei, R.J. and Sajja, P. (2012) ‘Mining gender affinity to social networking’, Int. J. Social Network Mining, Vol. 1, No. 2, pp.160–184.

Biographical notes: Rozita Jamili Oskouei received her BTech in Software Engineering from Tabriz Azad University, Iran in 2001 and MSC (IT) in Computer Science from Department of Computer Science, Jamia Hamdard University, New Delhi, India in 2007. She is currently a PhD scholar in the Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology (MNNIT), Allahabad, India. Her research interests include social network analysis and machine learning, web technologies, Semantic Web, education data mining, behaviour mining, web mining, information retrieval and ontology. She had been influential in
analysing the various affects of internet usages in students’ academic and personal behaviours. She had published 12 conference and journal articles and currently studying the relationship between internet usage and students medical problems, co-curricular and extra-curricular activities and their academic performance.

Phani Sajja received his BTech in Information Technology from Department of Information Technology, College of Engineering, GITAM, Andhra University, India in 2005 and MTech in Software Engineering from Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology (MNNIT), Allahabad, India in 2008. Currently, he is a Senior Software Engineer at Expert Software Consultants Ltd., Noida, India. His research interests include software engineering, distributed computing, web technologies, Semantic Web, information retrieval, data mining, and machine learning.

1 Introduction

Modern communication and computing technology have significantly influenced on daily activities and quality of life of individuals and organisations. A recent survey by Greengard (2011) has found 35% of the users check their email or Facebook account before getting out of bed in the morning. According to a survey by Lin and Yu (2008), the majority of internet users belong to the age group of 12–20 and for these users, online games, social network (SN) and e-mail are biggest attractions.

Different research efforts (Kraut et al., 1998, 2002; Lavin et al., 2000; Nie and Erbring, 2000; Stoll, 1995; Turkle, 1996; Widyanto and McMurran, 2004; Yao-Guo et al., 2006) have shown that this growing use of computing and communication technology has both good and bad effects. Kraut et al. (1998, 2002) claimed that greater use of the internet was associated with negative effects on individuals, such as diminishing social circle, and increasing depression and loneliness. In Figl et al. (2008), Figl et al. found those students who do not have a supporting network generally have fewer friends. Moreover, many quantitative studies confirmed that loneliness was associated with increased internet use (Lavin et al., 2000; Nie and Erbring, 2000; Stoll, 1995; Turkle, 1996; Widyanto and McMurran, 2004; Yao-Guo et al., 2006). High frequency and larger duration of internet use which interfere with daily life have been linked to a range of problems, including decreased psychosocial well-being, relationship breakdown (Turkle, 1996; Widyanto and McMurran, 2004; Yao-Guo et al., 2006). All these studies lead one to hypothesise that a controlled use of technology with a purpose, can result into increased productively and better quality of life. However, there is no confirmed survey about evaluation of positive or negative effects of internet usage on students’ personal and social behaviours in academic environments along with their academic performance.

These internet usage behaviours of young people, mostly students, raises many challenges to the academic researchers and makes an important issue for exploring students’ internet behaviours because they may be interested in the contents of undesirable websites, such as pornography. Revealing students’ behaviours might provide necessary knowledge for parents and teachers to guide students in their internet use and help them avoid the dangers.
Academic institutions in all around the world have made significant investment in computing and internet infrastructure to meet the following objectives:

1. to provide access to learning resources, which are available on World Wide Web
2. to significantly increase academic productivity of students and teachers and to enhance their learning experiences
3. to partially supplement the shortage of qualified teachers.

However, there is a growing perception in the community that these objectives are either not achieved or achieved marginally. One of the primary reasons cited for the above situation is that, these resources are dominantly used for non-academic (NAC) especially social networking purposes. Several academic administrators are advocated to blocking those websites, which do not directly relate to academic curriculum activities as a remedial major. Currently, there are no clear answers for the following questions:

- How much restriction is needed for NAC, and SN websites?
  - Is it necessary to block totally accessing of these websites?
  - Is it enough to make a limitation on SN websites access by each student?
  - Or no need for any limitation or blocking these categories of websites?

Answering the above questions requires more research efforts in different countries, different cultures and different levels of education and academic environments. Further, exploring effects of different factors such as student’s undertaken programme, branch, semester of study, gender and age on evaluating the internet usage behaviours of students in academic environments are necessary. For instance, female students, who have come from different social, cultural and demographic background and with the mass expansions of new technologies readily available, are exposed to something that is not present in their homeland. This sudden freedom of free flowing information and non-restricted use of communication through SN results in unbelievable changes in the behavioural patterns of these students.

In this investigation, an attempt is made for better clarifications on academic (AC) and non-academic (NAC) websites, which is taken into account for referred proposed websites classification scheme (Oskouei and Chaudhary, 2010; Oskouei, 2010a) and with the help of data mining techniques, a study is made to analyse students’ gender based internet usage in terms of category of visited websites and average time spent in each category especially NAC websites and SN websites per day in different periods of a semester including working days, holidays and examination days and relating those behaviours with their academic behaviours and activities and performance. Further, a survey is made based on analysing proxy server access log files. These access log files were collected for 24 months from 1st January 2008 to 29th December 2009.

This paper is organised in eight sections. Section 2 introduces basic concepts of web mining, behaviour mining and SN. Section 3 covers background and related works. Section 4 describes our data collection and pre-processing phase along with the website classification scheme. Section 5 presents users’ gender-based internet and SN websites usage behaviours. Section 6 demonstrates effects of different factors on students’ internet and SN websites usage behaviours. Section 7 presents a model for prediction of future academic performance [cumulative performance index (CPI)] of students based on their
internet and SN usage behaviours and finally Section 8 concludes the paper by summarising our achievement and depicting some interesting future issues.

2 Basic concepts

This paper is about analysing the SN websites usage behaviours of users with the help of analysis of proxy server access log files. These analysis are related to data, web and behaviour mining and SN analysis. This section provides a brief description of these concepts.

2.1 Data and web mining

Data mining is the process of extracting knowledge and interesting patterns from huge amount of data. It is the analysis of (often large) observational datasets to find unsuspected relationships and help to summarise the data in novel ways that are both understandable and useful to data owners (Srivastava et al., 2000; Srivastava, 2002). Web mining utilises the data mining techniques to automatically discover and extract information from web documents and services. Tanimoto (2007) outlined various suggestions of using this technology for educational purposes. Boyd and Ellison (2007) and Koohang (2004) clarified some other applications of web mining.

Web mining includes the following three sub-categories (Cooley et al., 1997; Kosala and Blockeel, 2000):

- **Web content mining** is concerned with the extraction of useful knowledge from the contents of web pages.
- **Web structure mining** is concerned with the application of data mining to the structure of the web graph.
- **Web usage mining (WUM)** discovers users’ interesting usage patterns by analysing web usage data stored in proxy server access log files or server log file. WUM (Bchner et al., 1999; Pierrakos et al., 2001) further consists of four basic stages: data collection, data pre-processing, pattern discovery and knowledge post-processing.

2.2 Behaviour mining

Identifying users’ behaviours on internet would be helpful for business owners and education environment’s administrators. Understanding internet users’ similar behaviours on business websites can be helpful for making better plans for increasing quality of productivity and marketing.

In academic environments, extracting users’ internet usage behaviours can be helpful for administrators for applying their favourite filtering or applications.

In WUM, different techniques for extracting users’ behaviours exist. For example:

- Analysing users behaviours by extracting the contents of the visited web pages (Butakov and Odinma, 2009; Oskouei, 2010b, 2010c; Suneetha and Krishnamoorthi, 2009).
User navigation behaviour mining (UNBM): studying the problems of extracting the interesting access patterns from user access sequences (UAS), which are usually used for web page recommendation (Xue et al., 2010; Iwazume et al., 1996; Berendt, 2000; Marques et al., 2004; Dong et al., 2006; Borzemski, 2007; Lin et al., 1999), fraud detection (Jalali et al., 2008) and predicting the future navigation behaviours of users on internet (Ling et al., 2007; Xu et al., 2007; Hao et al., 2001; Manavoglu et al., 2003; Yada et al., 2004; Adar et al., 2007; Zhou et al., 2005; Hu et al., 2007).

2.3 Social network

Srivastava (2008) defined a SN as a social structure of individuals, who are related (directly or indirectly to each other) based on a common relation of interest, e.g., friendship, trust, etc. SN websites provide different services and facilities for users. For example:

- Share favourite files (with various formats) with others such as photos, music, movies, etc.
- Creating an online community of internet users, researchgate.net is a professional communication SN website.
- Articulate a list of other users with whom they share a connection, view and traverse their list of connections those made by others within the system. The nature and nomenclature of these connections may vary from website to website. [A concise definition as given by Boyd and Ellison (2007)].
- Individuals can set up an online profile describing his/her interests, and adds links to other profiles. Generally, users are able to post personal information, including photographs, videos and blog entries. Some of the popular SN websites are Facebook.com and Orkut.com.

Therefore, social networks operate on many levels, from families up to the level of nations, and play a critical role in determining the way problems are solved, organisations are run, and the degree to which individuals succeed in achieving their goals.

SN analysis is the study of social networks to understand their structure and behaviour. The study of social networks has received significant interest from researchers in various domains such as psychology, philosophy, education, and lately computer science – particularly in the field of artificial intelligence.

3 Related works

Several studies have been conducted in order to investigate the increasing of number of internet users, and their time spent on the internet daily. The study made by Rao and Flurry (2011) in June 2011 shows that users spent an average of 74 minutes on the web, compared to 70 minutes on the web daily in December of 2010. For the larger web brands, users spent an average of one hour 53 minutes a month on Google, three hours eight minutes on Yahoo and five hours 24 minutes on Facebook. These usage study compliments other reports (Warren, 2009; Porterfield, 2010) that 86% of internet users visited social networking websites.
Discovering usage patterns from web data is one of the primary purposes of WUM. Several studies (Yen and Lee, 2001; Chevalier et al., 2003; Udechukwu et al., 2004; Lu et al., 2005; Nina et al., 2009; Maheswara et al., 2010) have been made to extract web navigational patterns of users, including frequent access patterns, sequential access patterns to model users’ behaviours. Nina et al. (2009) analysed the web access log files and by applying pattern discovery methods, discovered users location or country by looking for the domain name and then used this information for assuming the most active countries’ visited website and for providing the useful information relevant to that country.

Several researchers studied gender-based effects of information technology in particular on education (Butakov and Odinma, 2009; Oskouei, 2010b, 2010c; Suneetha and Krishnamoorthi, 2009) and (Jarrell, 1994; Rasmussen, 1998; Stevens, 1984; Stepanikova et al., 2010; Erdoan, 2008; Stromquist, 1989; Abbe and Momodu, 1999; Welty, 1996; Sanders, 2005; Huynh et al., 2006; Kim and Chang, 2007; Cox and Fisher, 2007; Al-Radaideh et al., 2006). Stromquist (1989) stated that in Sudan 43% of graduating females over a decade ago were in liberal arts, in contrast to 14% in medicine and 14% in sciences. It is also said that, in a very few circumstances do women select agricultural or engineering fields. A study based on 551 university students and graduates in Tanzania in 1979 found that only 1% of the female students compared with 16% of the males were enrolled in higher technical-related colleges. Further, according to the study of Abbe and Momodu (1999), women’s education positively correlates with several national and international goals, some of which include economic productivity, social development, social equity and sustainable development. Braten and Stromso (2006) examined the contribution of epistemological beliefs, individual interest, and gender to self-reports of internet-based learning activities and concluded that, males reported higher levels of participation in internet-based communication activities than females, and females reported higher levels of strategy use when learning from conventional printed texts than males. Kim and Chang (2007) study has concluded that computer usage have differential effects on academic performance of users from the immigrant and gender groups. It advocated adaptive design of multimedia contents for different groups of students addressing their specific needs. Lin and Yu (2008) explored gender differences in adolescent internet accessibility, motives for use, and online activities in Taiwan. Their results revealed for media such as newspapers, radio, and magazines there was significant difference, with boys tended to view these media as more important than girls did. Butakov and Odinma (2009) studied the internet usage behaviour of students at the American University of Nigeria and concluded that the internet resources were not used for academic related requirements and were mostly used for entertainment by both genders (female 60%, male 70%). Further, female users spent more time on particular websites, especially SN websites. Erdöan (2008) studied the internet attitudes and loneliness of students on Turkish adolescents and reported that Turkish adolescents’ loneliness was associated with both increased internet usage and internet attitudes.

Some of the researchers (Al-Radaideh et al., 2006; Braten and Stromso, 2006; Awais et al., 2008; E-Halees, 2008) studied the impacts of internet usages on students’ academic performances. Al-Radaideh et al. (2006) proposed a model to predict the performance of students in the final examination of C++ course based on their performance in the mid-term examination, their attendance and their activities on C++ webpage, answering questions and assignments of that webpage. Awais et al. (2008) explored the impact of internet usage on students’ academic performance. For this analysis, data were collected
from six different universities with a questionnaire methodology. Correlation and regression techniques were used to find out the strength and direction of the relationship. Their results showed 49% of students agreed their internet skill were high whereas 18.2% of these internet users had CGPA lesser than 2.5. El-Halees (2008) extracted learning behaviours of students by analysing students’ data from database management system course and their usage of e-learning websites. They concluded that e-learning facilities such as e-resources used by students, exercises and assignment hardly affected the final grade of students. The research study by Liccardi et al. (2007) revealed the role of social networks in students’ learning experiences. This research concluded that social networking plays a positive role in students’ learning experiences.

4 Data collection, pre-processing and website classification scheme

4.1 Data collection and pre-processing

For this analysis, we collected proxy server’s access log files from the computer center of Motilal Nehru National Institute of Technology Allahabad, India which runs on ‘Squid 3.0 stable’. The data collection period was 24 continuous months, which includes four academic semesters. These log files are used for debugging, user and site profiling and measurement of resource utilisation. The computer center also provided user identification information, which includes user-id, full name and department in one text file. Further, users’ academic information are collected from the dean (academic affairs) office in one Microsoft excel file. The fields which are included in this excel file are: registration-no., full-name, programme, branch, semester, gender, CPI and semester performance index (SPI).

Each record of the access log file has 11 fields for each visited web page by one individual user. For our analysis, we have selected three fields out of these 11 fields of each record. These three fields are: user-id, date and time of connection and the URL of the visited web page.

For statistical analysis, we used Rapid-Miner and R-Miner tools. The records with inconsistent or missing values are removed during the pre-processing step. An algorithm is developed and implemented to compute the time spent by users on different categories of websites along with the total time spent per day by each user, by using time stamp of the access log files. Minute is used as a unit of time duration to measure the time spent by one individual user on each website during a day. This unit of measurement implies that all durations less than a minute have been rounded to one minute. The proposed algorithm also computes the average total time spent by each user per day and average time spent by him/her on different category of visited websites.

4.2 Website classification scheme

Several website classification schemes are proposed based on contents and structure of the websites. Most of these classification schemes are public. The classification scheme proposed by ODP (dmoz, 2011) is most popular than other classification schemes and is mostly used by search engines. Whenever the research purpose is concerned with a special area, identifying a new classification scheme is necessary. Since our investigation is related to analysing data from education area and according to the goals of academic
environments, the ODP categories to which websites visited by students belong, do not relate to activities of these academic environments. Therefore, we need to have concepts in the classification scheme which explicitly relate to activities of students in residential academic institutions. Students’ activities are generally classified as academic (AC) and non-academic (NAC) activities. These concepts have been introduced either as generalisation or specialisation of ODP concepts in our proposed classification scheme.

**Figure 1** Websites classification scheme (see online version for colours)

Websites which are used primarily related to curricular activities are classified as academic websites (AC) and those websites which are used for other purposes, which are not related to curricular activities, are classified as non-academic (NAC) websites. Further, academic-websites (AC) must be cover activities related to education such as Conferences of IEEE, ACM, Springer, etc., e-books, and colleges or institutes websites, programming, research organisation websites, or dictionary and free translators, etc., non-academic (NAC) category might cover different activities which students are interested, such as online-game, music, news, etc.
Website classification scheme used for our analysis is given in Figure 1. Websites classification scheme in Figure 1 is based on the analysis of approximately 300,000 websites visited by students for 24 months. Further, this classification scheme is based on generalisation or specialisation of the ODP classification. More details of the proposed classification scheme are given as:

- **Academic websites (AC):** is generalisation of reference and science category of ODP which are represented in terms of general sub-category of academic-websites. The other extended sub-category which is named as professional and it contains commercial and non-commercial sub-categories. Some of the sub-categories in commercial and non-commercial are the same, for example, e-book websites on the internet, which are available for free as well as for sale.

- **Non-academic websites (NAC) includes:**
  - **Portal:** these websites provide many types of services such as social networking (SN) [e-mail, chat, etc.], entertainment (EN) [video, music, photo, news, online games, etc.], business [e-shopping]. Yahoo.com is one of such portal.
  - **Entertainment (EN):** includes those websites which provide entertainment facilities for users
    1. media: includes online-TV and radio, video, music, actor and actress photos, etc.; [http://www.iranonlinetv.com](http://www.iranonlinetv.com)
    2. online game: [http://zumadeluxeonline.info/](http://zumadeluxeonline.info/)
    3. news: [http://www.bbc.co.uk](http://www.bbc.co.uk)
    6. undesirable websites: including pornography and drug websites
  - **Personal websites:** this category of websites are slightly different from blogs. [http://www.hawking.org.uk/](http://www.hawking.org.uk/)
  - **Government websites:** includes websites which are providing government services for people including railways, passport offices, embassies websites, etc.
  - **Social networks:** based on the usage purposes or services provided by these websites we further classified three sub-categories for these websites:
    1. Advanced social networking websites: the purpose of creating these websites are to establish online social and communication networks between people around the world. Facebook.com and Orkut.com are the two big such websites.
    2. Blog: blog is a type of website which is maintained by an individual or a group with changeable entries of information, description of events, graphics or video. Entries are commonly displayed in a certain chronological manner. [http://www.blogger.com](http://www.blogger.com) and [http://www.blogfa.com](http://www.blogfa.com) are the example blog sites.
    3. Special SN websites: including
• **Sharing websites:** those websites which provide facilities to share multimedia content including photos, music, video, text, etc. (e.g., Photobucket.com, youtube.com)

• **Online communities:** http://www.researchgate.net.

• **Business:** includes those websites which are designed for marketing or shopping purposes

  1. advertising: http://www.advertising.com/
  2. e-shopping: http://www.ipkart.com
  3. job-agency: http://www.nukri.com/
  4. company: http://www.tcs.com/
  6. e-commerce: http://www.forex.com/
  7. travel agency: http://www.makemytrip.com/

It is interesting to note that, some of the websites can be categorised in more than one category. For example, undesirable websites can be categorised both under business (not free) or entertainment (free of cost for visiting or downloading contents) categories.

### 5 Differential gender-based internet and social networking websites usage

This section provides the analysis results of students’ internet connection and SN websites visited during 24 months. In this section, gender-based differential internet connection along with social networking websites usage behaviours are also studied.

Several research efforts (Huynh et al., 2006; Kim and Chang, 2007; Cox and Fisher, 2007) have been made for exploring the effects of gender behaviours and internet usages of students. Zuga (1999) finds that women are technologists and have always contributed to the making of their environment. She examines the process of involving girls and women in the courses and the necessity for restructuring the subject matter and incorporating feminist ideals into the pedagogy. Cox and Fisher (2007) formed project groups that were based on sex of the group members, thus by creating one female group and other male groups. They explored all software used by each group and asked them to write their feelings for using those software and online courses and communication tools available on internet. The overview reported that users in different ages having various behaviours and feelings are even using similar software. Huyhn et al. (2006) focused gender-based group study on perspective and experiences of female management students and their instructors in an online learning environment that uses a computer supported collaborative learning (CSCL) system to facilitate student-to-student interactions. They found that females have a greater involvement in designing and assessing various features of CSCL.

Our study is focused on analysing in depth the usage patterns of websites belonging to academic (AC) and non-academic (NAC) especially SN websites.

Our primary results show that, out of 300,000 websites visited by students during four academic semesters, approximately 36% of the total usage was belonging to AC websites and 64% was related to NAC websites. Therefore, we put a special attention in this
investigation to study the usage patterns of NAC and SN websites. Since SN is one of the important and most popular visited categories, all sub-categories of these SN websites along with the effects of gender, semester, branch and programme wise on SN usage patterns of students are also analysed.

5.1 Analysing students’ internet usage behaviours

In most of the countries female students are in minority in engineering colleges. In our study, the population of female students was approximately 14%. Therefore, this study is concerned on extracting gender-based differences on internet usage behaviours of students in engineering colleges. These analysis results can be helpful for instructors, course coordinators and parents for their plans and decisions in the future.

Our analysis in this section is based on exploring different aspects of students’ internet usage behaviours such as: average time spent on internet and especially on SN websites, average number of visited web pages per day, percentage of different categories of visited websites per day, relationship between users’ academic activities in terms of CPI and their average time spent per day along with proposing a method for predicting students’ final examination results based on their SN usage behaviours. Total undertaken student population in this study was 5210 during four continuous academic semesters. These students were undertaking different degree programmes such as BTech, MTech, MSc, MBA, MCA, PhD. Therefore, the next observation would be exploring relationship between students’ undertaken programme (such as BTech, MTech, MSc, etc.) or semesters and their internet or SN usage behaviours.

5.1.1 Differential internet usage behaviour based on time, and number of visited web pages

Since most of the engineering colleges are providing internet facilities for students and faculties, exploring internet usage behaviours in terms of category of visited websites and average time spent by each individual user would be our primary analysis for the purpose of identifying differential usage behaviours among students.

This research is based on the analysis of the access log files, which are collected during 24 months from proxy server of Motilal Nehru National Institute of Technology (MNNIT) Allahabad India, which has a decentralised computing environment. Each academic and administrative department/section has their own computing facilities in addition to the computer center, which is a central facility. The institute has approximately 1,000 computing nodes distributed all over the campus and connected through optical fibre backbone including hostels and residential areas. The computer center houses approximately 300 nodes and operates from 07:00 am to 11:00 pm for 365 days. Computing nodes within the hostels are of the student’s and are not included in the above count. Internet connectivity is provided through 42 Mbps leased lines.

Since unlimited access to internet on working days and holidays during each semester is provided, therefore our first observation would be exploring the average time spent by students on internet per day in different periods of a semester. Further, in the next observation, we would be looking for any changes on these usage patterns and time spent during examination weeks and the day before examination for identifying effects of examination tensions or stress on students’ internet usage behaviours. Table 1 shows the average time spent per day by female and male students during a semester.
Table 1  Gender-based internet usage pattern (in terms of time-spent on internet)

<table>
<thead>
<tr>
<th>Group description</th>
<th>Usage (in %)</th>
<th>Average time (mins)</th>
<th>Average visited web pages</th>
<th>Min time-spent</th>
<th>Max time-spent</th>
</tr>
</thead>
<tbody>
<tr>
<td>All users</td>
<td>78</td>
<td>104</td>
<td>608</td>
<td>1</td>
<td>708</td>
</tr>
<tr>
<td>Female</td>
<td>24</td>
<td>204</td>
<td>351</td>
<td>1</td>
<td>351</td>
</tr>
<tr>
<td>Male</td>
<td>76</td>
<td>93</td>
<td>681</td>
<td>1</td>
<td>708</td>
</tr>
</tbody>
</table>

Second column of this table shows the population of users. From these results, 78% (4,055) of total undertaken students (5,210) had connected to internet during our study. In other words, 22% (1,155) of the users never connected to internet or connected with other users’ user-id. Since for each user one unique user-id is provided by the computer center, whenever a user connects to internet, for each visited web page by this user, one record will be added in the access log file. Based on our results, some of the users connected with other’s user-id. Since we have IP-address of each user, we are able to identify some of these misuse. However, we will not discuss these issues in this paper.

Out of 4,055 users, 24% of users were female and 76% of users were male. Average time spent per day by male users was 93 minutes and for female users it was 204 minutes. This means an average time spent by female users was more than male users but number of visited web pages by female users was less than male users. In other words, female users spent more time for visiting each web page.

The next observation is related to the category of visited websites by students per day. The main purpose of this analysis is to figure out the percentage of academic websites which are visited by students during one academic semester along with revealing the category of popular websites. Table 2 presents the percentage of students who visited AC and NAC categories of websites along with average time spent per day on each category.

Table 2  Identifying the category of visited websites by students

<table>
<thead>
<tr>
<th>Website category</th>
<th>Group description</th>
<th>Usage (in %)</th>
<th>Average time (mins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic websites (AC)</td>
<td>All users</td>
<td>31</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>35</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>28</td>
<td>23</td>
</tr>
<tr>
<td>Non-academic (NAC)</td>
<td>All users</td>
<td>69</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>65</td>
<td>154</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>72</td>
<td>84</td>
</tr>
</tbody>
</table>

It is evident from Table 2 that,

- the majority of internet usages by students belongs to NAC category of websites
- further, the average time spent per day by students on NAC category of websites was more than AC category of websites
- 35% of female and 28% of male students visited AC category of websites but average time spent by male users on AC websites was more than female users
- time spent on NAC websites by female students was more than male students.
Therefore, after recognising this fact that NAC category of websites is more time consuming for students during a day, it would be interesting to explore the sub-categories of NAC websites in which the students spent more time. Discovering more popular NAC sub-categories, which are visited by most of the students, would be useful for future actions or decisions which might be taken by academic advisors, administrators or by parents for different purposes including improving students’ academic activities. Distribution of users based on their visited sub-categories of NAC websites are given in Table 3.

**Table 3 Extracting students’ usage of non-academic websites**

<table>
<thead>
<tr>
<th>Usage description</th>
<th>Group description</th>
<th>Usage (in%)</th>
<th>Average time (mins)</th>
<th>Max time (mins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social networks websites (SN)</td>
<td>Female</td>
<td>37</td>
<td>125</td>
<td>705</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>27</td>
<td>70</td>
<td>752</td>
</tr>
<tr>
<td>Entertainment websites (EN)</td>
<td>Female</td>
<td>25</td>
<td>32</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>31</td>
<td>106</td>
<td>280</td>
</tr>
<tr>
<td>Business</td>
<td>Female</td>
<td>1</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>10</td>
<td>12</td>
<td>41</td>
</tr>
<tr>
<td>Other subcategories</td>
<td>Female</td>
<td>2</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>4</td>
<td>10</td>
<td>55</td>
</tr>
</tbody>
</table>

From Table 3, it is evident that:

- On average SN websites are most popular visited websites by students. In other words, 37% of female and 27% of male users spent on average 125 and 70 minutes times per day in these websites respectively.

- The second popular websites visited belong to EN category. Majority of EN visitors belongs to male users and they spent on average 106 minutes per day in these websites.

- There are gender-based differences on the category of visited websites. The majority of female usages belongs to SN websites whereas, for male users, the most favourite visited websites were EN websites. Further, the usage percentage of business category of websites by female is different from male users.

This investigation is related to analysing SN websites usages on students’ activities and performances, therefore in the remaining part of this section we will concentrate on analysing in more details these usages.

### 5.1.2 Gender-based differential behaviours on SN websites

Table 4 presents more details of SN usage by female and male users. In Figure 1, we defined three sub-categories for SN websites; therefore it would be interesting if we figure out which sub-category is the most popular for students and extract the average time spent by users in those sub-categories.
Table 4  Users’ gender-based usage pattern of SN

<table>
<thead>
<tr>
<th>Usage description</th>
<th>Group description</th>
<th>Usage (in %)</th>
<th>Average time (mins)</th>
<th>Min time (mins)</th>
<th>Max time (mins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced SN</td>
<td>Female</td>
<td>37</td>
<td>240</td>
<td>5</td>
<td>705</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>57</td>
<td>129</td>
<td>3</td>
<td>752</td>
</tr>
<tr>
<td>Shared SN</td>
<td>Female</td>
<td>40</td>
<td>63</td>
<td>1</td>
<td>537</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>29</td>
<td>45</td>
<td>1</td>
<td>182</td>
</tr>
<tr>
<td>Blog</td>
<td>Female</td>
<td>23</td>
<td>9</td>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>14</td>
<td>15</td>
<td>1</td>
<td>35</td>
</tr>
</tbody>
</table>

The results of above table reveal a gender-based different usage behaviours of SN sub-categories. The results of this table are summarised as:

- The majority of advanced SN usages belongs to male users. 57% of total male users visited advance SN sub-category and spent an average of 129 minutes time per day on this sub-category of websites, whereas 37% of female SN users visited advanced SN sub-category and they spent an average 240 minutes time per day in this sub-category. On the other hand, the majority of advanced SN sub-category usage belongs to male users, but female users on average spent more time than male users in these websites.

- The majority of sub-category of sharing SN websites usages belong to female users (approximately 40%) and they spent more time than male users in these sub-category.

- The majority of blog usage belongs to female users, whereas on average more time spent in this sub-category belongs to male users. On the other hand, number of male users, who have blogged or visited others’ blogs were lesser than female users, but at the same time, these male users spent more time for reading other’s blogs or editing their own blogs in comparison to female users.

It is interesting to note that male users mostly visit blog sub-category in the morning time approximately between 6–8 a.m., whereas female users visit this sub-category in the evening time between 6–8 p.m. Further, sharing SN sub-category is visited by female users between 3–5 p.m., whereas male users visited this sub-category between 1–3 a.m. However, there are no differences between female and male users regarding usage timings of advanced SN sub-category during a day because most of the users from both genders are online or at least visit these websites at the same time.

6  Analysing the effects of different factors on the internet and SN usage behaviours

6.1  Effects of examination stress on SN usage behaviours

First, we explore all the changes, which probably happen during the examination periods on female and male users’ usage behaviours of SN websites. Figure 2 represents the
average number of users connected to internet during 24 hours of the day in different periods of a semester, including normal days of a semester, the day before examination and during examination weeks. The vertical axis in this figure represents the average number of users, and the horizontal axis represents 24 hours of the day. It is evident from this figure that the usage pattern of users is almost similar during different periods of a semester. In other words, the maximum number of users are connected to internet at 4 a.m. and between 13–17 p.m. Another observation is that the majority of SN users connected during the final examination weeks, whereas minority of SN users connected during a day before examinations.

Figure 2  Population of users on during 24 hours per day on different periods of a semester
(see online version for colours)

The average number of visited web pages during different periods of a semester is presented in Figure 3. The vertical axis of this figure represents the average number of visited web pages and the horizontal axis represents 24 hours of the day.

This figure shows that the average maximum number of visited web pages by students belongs to examination weeks, whereas the minimum number of visited web pages belongs to a day before examination weeks. Therefore, based on the results of Figure 2 and Figure 3, it is evident that the minimum number of SN websites visitors and the minimum number of web pages visited belong to a day before examination. This might happen due to stress or tension of examination.

It is interesting to note that based on our analysis during examination weeks, the maximum time spent on the internet by female users belongs to EN/undesirable and SN websites, whereas the maximum time spent by male users during examination weeks belongs to EN/online-game, AC/science, AC/professional/free-e-book or programming websites categories. More details about these behaviours can be found in Oskouei and Chaudhary (2010).
6.2 Effects of users’ undertaken branch and semesters on their SN usage behaviours

The next observation is related to analysing the effects of student’s undertaken programme such as BTech, MTech, MSc, MCA, etc., on their SN usage behaviours. The gender-based population of students for each one of degree programmes is given in Figure 4. The vertical axis in this figure shows the percentage of students’ population, and the horizontal axis shows different programmes. This figure shows that male students are the majority in all degree programmes. Further, female students are in minority on MTech degree programme compare with other programmes.

The population of SN users based on their gender and undertaken programme is given in Figure 5. The vertical axis shows the percentage of users and the horizontal axis shows different degree programmes.

From this figure, for both female and male users, the maximum number of SN users belongs to BTech degree programme, whereas the minimum percentage of both male and female users belong to PhD and MBA degree programme respectively.

It is interesting to note that, based on our previous research, most of the internet users have undertaken BTech degree programme in different departments. Moreover, students of some branches such as mathematics or chemistry never connected to the internet. The maximum population of SN users whose undertaken.
BTech degree programme belongs to IT, computer science and engineering, electronics and electrical departments. In other words, there was a relationship between undertaken degree programme, especially branch with users’ SN websites visitbehaviours.

The other interesting point from our analysis was related to fresher students’ internet and SN usage behaviour differences with other students. The fresher students refer to the students of first year.

The distribution of BTech students along with the visited SN websites during an academic semester is given in Figure 6. The vertical axis shows the percentage-based distribution of SN users, and the horizontal axis shows the different academic semesters with which users are undertaken those semesters.
There are small differences on SN usage behaviours of female and male users as it can be observed from Figure 6. The maximum number of SN female users belongs to the first year (the first and the second semesters) and the minimum number of SN female users belongs to the final year (the seventh and the eighth semesters). In other words, the majority of SN female users belongs to female students of the first year. These students who came from different states with various cultural and economic backgrounds mostly connect to SN websites to overcome the loneliness and to adapt to new social and academic environments. However, female students of the final year (fourth year) are in minority compared with other academic years regarding SN usage. The majority of male SN users are in the fifth and the sixth semesters (third year) and minority of users belongs to the first and second semester (first year). In other words, female and male users had different behaviours when starting life in a new environment and for adapting to new social and academic environment, and these deferential behaviours have influenced their SN usage.

6.3 Academic activity and SN usage behaviour

With the growing usage of SN websites, the main question for the academic advisors, professors and parents is what are the effects of SN websites usages on the students’ academic activities and performances. Most of the parents worry when they see every day their children spent more or less time on the internet, especially for SN communication. This section presents the details of students’ academic performance relationship with their SN websites usage along with their time spent on those websites. Figure 7 presents the average CPI of female and male SN users. From this figure, the average CPI of SN users was 7.8 and 7.2 for female and male respectively.

Since CPI scale is in the range of 0 to 10, students with CPI less than five are academically weak students. Figure 8 presents the average time spent by female and male users whose visited websites belongs to SN category. The vertical axis represents the average time spent in minutes per day in SN websites, and the horizontal axis represents the different ranges of CPI.
It is evident in the figure that there is a gender-based difference between female and male users regarding their CPI and time spent on the internet. Female users with CPI < 5 had spent minimum time on SN websites, whereas the maximum time spent by female on SN websites belongs to users with 8 ≤ CPI < 9. These results are not valid for male users’ SN usage. In other words, the maximum time spent on SN by male users belongs to those users with CPI < 5 and the minimum time spent belongs to those are in 9 ≤ CPI ≤ 10 range. On the other hand, academically the weak students have different behaviours regarding time spent on SN websites. Female weak students spent less time than the other females on SN websites, whereas it is completely reversed for male users.
7 A method for predicting students’ CPI based on their SN usages

In this section, we follow students’ categorisation based on their academic performance which is used by most of the universities and colleges in India. This categorisation scheme has four categories based on CPI. Table 5 gives the details of these categories. The category A indicates academically excellent students, B indicates academically good students, C indicates academically normal students and D indicates academically weak students.

Table 5 Categorisation of students based on their academic performance (CPI)

<table>
<thead>
<tr>
<th>Group no.</th>
<th>Group name</th>
<th>CPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Excellent students (A)</td>
<td>CPI ≥ 8.5</td>
</tr>
<tr>
<td>2</td>
<td>Good students (B)</td>
<td>7 ≤ CPI &lt; 8.5</td>
</tr>
<tr>
<td>3</td>
<td>Normal students (C)</td>
<td>5 ≤ CPI &lt; 7</td>
</tr>
<tr>
<td>4</td>
<td>Weak students (D)</td>
<td>CPI &lt; 5</td>
</tr>
</tbody>
</table>

We used k-means clustering method for grouping SN users based on their time spent. Table 6 shows the centroid and density of each cluster.

Table 6 Clustering SN users based on their time spent on these websites per day by k-mean

<table>
<thead>
<tr>
<th>Cluster</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centroid</td>
<td>5</td>
<td>22</td>
<td>68</td>
<td>105</td>
<td>139</td>
<td>251</td>
<td>479</td>
</tr>
<tr>
<td>Density</td>
<td>16%</td>
<td>13%</td>
<td>22%</td>
<td>9%</td>
<td>15%</td>
<td>7%</td>
<td>17%</td>
</tr>
</tbody>
</table>

The main purpose of our analysis is concerned on predicting students’ future academic performances in terms of CPI for helping academic advisors, course coordinators and parents by recognising the students at risk before examination in the context of possibilities of their failure in coming semesters based on their internet access and SN usage behaviours. This prediction would be helpful for improving academic results of students in each academic environment.

We applied decision tree for predicting future examination results. This decision tree would be useful for predicting CPI of final examination by using users’ gender, average time spent on internet and especially on SN websites per day, their undertaken semester, degree programme. With the help of these variables, our proposed decision tree will predict each individual student’s category (A, B, C or D) in the coming semester examination results. Part of this decision tree is given in Figure 9.

This part of the decision tree shows that if time spent on SN websites is in the range of 180 < time < 240 minutes per day and students are in second semester, without any gender-based differences, he/she will be in the group of academically excellent students (A). But if the student is not in second semester, then there are gender-based differences. Therefore, if the student is male, and he has undertaken BTech degree programme and his semester is less than or equal to 6, then he will be in the group of academically excellent students (A), but if he is in seventh or eighth semester, then he will be in the group of weak students (D). Further, if the user is female and her degree programme is MCA, then she will be in academically normal students (C), but if she has undertaken BTech degree programme, and semester less than 8, then she will belong to students with good academic performance (B) otherwise she will belong to academically normal students.
On the other hand, there is no possibility for female students with time spent more than 180 and less than 240 to become academically excellent students (A). These results can be validated with Figure 8.

**Figure 9** Part of prediction made by decision tree (see online version for colours)

This decision tree has been tested for two continuous semesters and the accuracy of 81% for predicting female users’ CPI and 77% for predicting male users were obtained.

### 8 Conclusions

This investigation presents a comprehensive study of the impact of students’ internet usage behaviours on the academic performance in a country such as India with different cultures, languages and customs. The gender-based differences between students’ internet and SN usage behaviours in terms of average time spent and average number of visited web pages per day on SN category of websites along with different sub-categories of SN are studied. Further, the effects of different factors such as the students undertaken degree programmes, semesters are explored. Based on the results of our study, academically weak female students spent lesser time on SN websites whereas it was not true for weak male students. On the other hand, this research shows that using SN websites have no direct negative effects on students’ academic performance.

This research revealed two important factors in evaluating the effects of SN usages on students’ academic performance which are: average time spent per day on these websites and gender differences. Our overall results show that SN websites usage with some limitation on total access time per day might be suggested for some students, especially
for male students. We believe that the results which are obtained in this study might be different in other countries especially in Arabian countries or European countries, where the facilities provided for female students are different.

References


Mining gender affinity to social networking


