World Applied Sciences Journal 12 (Special Issue on Computer Applications & Knowledge Management): 47-54, 2011
ISSN 1818-4952

Finding Knowledge in Students Social Network

[1]Nur Shaqifah Mohd Sani, [1]Shakirah Mohd Taib, [2]Kamaruzaman Jusoff and [1]Ainol Rahmah Shazi

[1]Department of Computer and Information Sciences, Universiti Teknologi PETRONAS,
Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia
[2]Department of Forest Production, Faculty of Forestry,
Universiti Putra Malaysia, 43400 Sertdang, Selangor, Malaysia

Abstract: Social networking has been one of the widely used approaches in the communication technology movement. It is become a new trend of getting connected to other people and definitely it stores huge data including user activities and their shared materials. Many have seen the importance of collecting data for future benefits. In recent years, many companies have successfully analyzed their customer behaviour using various data mining techniques. One of the latest applications of data mining is in social network sites or environments. The objective of this paper is to present the analysis of social network user behaviour using clustering technique and centrality coefficient on university students' involvement. The result of the analysis is then validated with a questionnaire-based personality test. The study discovers the patterns of students' participation in social networking can be related to their personal behaviour that reflected by their characteristic and online activities. The analysis extends the research on promoting dynamic study culture at the higher learning institutions through online social network.

Key words: Social network · Clustering · Centrality coefficient · Data mining

INTRODUCTION

Organizations have started to realize the advantage of social network analysis and data mining in their businesses. Therefore these approaches now have been applied in many relevant fields. Business organizations and companies apply Social Network Analysis (SNA) and data mining mainly to identify their current customer's behavior and trends. The information gathered are used in the extrapolation of customer behavior for example in the prediction of how customer will react to their products. Thus the analysis is very useful to plan the next best steps that can be taken to improve their businesses.

The extent of the number of Social Network Sites (SNS) users has brought much interest from researchers to discover more knowledge from its hidden potentials. Millions of users integrate the SNS such as Facebook and Friendster into their daily practice as a medium of communication [1]. According to Pempek *et al*., SNS is the preferred medium for students to construct their profiles and engage in activities that reflect their identity [2]. The recent social network boom has significantly influenced the lifestyle of the students. The tendency of students to spend their most time on social networking sites has given rise to a new communication culture.

The objective of this paper is to propose that based on the current social networking trend amongst students, SNS is poised to be a main channel for knowledge creation and dissemination among this demographic group in which it is supported by the analysis of members' data obtained through mining the activities of a certain student group on a specific SNS.

Social Network Sites: SNS is defined as a web-based application that introduces users to create a public or controlled profile in a large and systematic system as well as sharing similar interest within a shared network [1]. Beginning from their infant years on the basis of documents sharing, to finding old classmates online, to its ultimate boom through applications such as Friendster, Facebook, Myspace and many others, SNS have been part of the current generation's agenda [3].

SNS require the user to sign up for services in which certain important and basic information need to be given to the system. This is the basic way of segmenting the user in the system. Users are then creating a "public" profile that can be viewed by others in the similar network. Privacy issues have been rather overlooked due to the euphoria of the ease of connection and interaction these SNS support. Dwyer *et al.* has perform a comparison

Corresponding Author: Shakirah Mohd Taib, Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia, E-mail: shakita@petronas.com.my.

between Facebook and MySpace, found that in online interactions, trust is somewhat less important in the creation of relationships as compared to in the offline world [4].

Once a user starts joining a network, the user starts to create extended relationships. The more friends they have, the more updates they will get [5]. The basic rule of SNS is that the user's level of activity within the network has an effect on the size of the user's personal network [1].

Social network activities are however never static. The movement of the network has created changes in SNS [6]. For an organization such as a university, this factor can lead to a great discovery that can be manipulated especially finding hidden behaviors of students.

Social Network Analysis: The process of finding hidden elements or properties in a social network is called as Social Network Analysis (SNA) [7]. Many researchers conduct the analysis with the interest from various aspects. According to Brandes, social network consists of relationship between nodes or actors in a network where it is theoretically modeled as graph, $G = (V, E)$, V is the actors and E as relations or potential attributes [8].

Through SNA, a particular organization can have a deeper understanding on the activities or events that happen around its community. For example, in a college circle, the network can be analyzed through the relationship between the students to students and between the students to their lecturers. However, other external factors may involve like getting interactions from other staffs of various departments. This leads to a more complex and extended version of relationship if details of the social network are derived continuously.

According to [9], AGNA is a platform-independent framework that analyzes social network by computing two categories of parameters; centrality coefficient and sociometric indexes. Centrality coefficient is computed to find the degree of access to information of an actor in a social network. Meanwhile the sociometric indexes measure the level of communicational activity of a specific actor. Another type of analysis that can be done using AGNA is the distance related analysis that is based on the concept of geodesic distance. It comprises of measuring the length of the shortest path between two nodes (e.g. shortest possible path from node A to node B) where length is calculated based on number of edges connected to a node. If all shortest paths connect to one node, then, that node can be considered as very central [10]. A link distance that based on geodesic concept has many advantages such as computational efficiency and is easily visualized [11].

Social Network Mining: Social Network Mining (SNM) is the implementation of data mining on social network sites. Information posted on profiles is mined to identify hidden data that can be manipulated. For example, mining keywords posted by users on profiles or status can lead researchers to identify the users' activities and whereabouts.

SNM has been used to find the hidden patterns or important information over large data size. This includes database consisting associated objects like people, events, organizations and other possible objects that may related to each other [12]. SNM is also applicable to other organizations. For example, in government sectors, SNM can be used in intelligence activities. This can be considered as a part of defensive approaches where a country can create an expectation on what kind of possible threats that might strike the country. For example, in United States, the concern over terrorist attacks has increased the usage of network mining in its defense system [13].

One of the advantages of mining the social network is that it allows people or organization to improve their performance. This is true in its ability of discovering new knowledge, technology and capabilities through its great models and techniques. However, SNM can also create a chaotic situation like the issue of violating privacy and confidential incursion [14]. Therefore, this technology can be one of threats to the data mart which maybe too loose to allow data to be mined and analyzed.

Related Works: Several researchers have initiated the motivation to implement data mining in higher education. According to [15], during his initial study of data mining implementation in higher education, the most valuable features of data mining technique were clustering and prediction. He identified how clustering can be used to analyze student's characteristics while prediction can be applied to track any possible outcomes from the students such as their achievements and involvements in classes.

The usage of decision tree algorithm like classification and regression tree (C and RT) has increased the capability of data mining ahead of traditional analytical techniques [15]. Another example is to have a system that can assist students to make decisions on academic programs based on analysis using

data mining technique [16]. The system is meant to help students in selecting courses when they do their course enrollment. Based on data collected through experiences and best practices of previous students, the system allows the prospective students to identify which course is suggested for them.

MATERIALS AND METHODS

Based on the research methodology used by Cain, the method used considered observation of actors, activities and relationship in an organization. The research method is divided into four parts. (a) Identifying the context of study, (b) Incorporate and validate (c) Observe and model organization and (d) Simulate, analyze and interpret the result [17].

In order to identify research context, the process of finding and understanding the best information, data mining technologies and approaches that can be conducted within the research scope are determined. A detailed study was done on how data mining and SNA could be applied in higher institution environment. In this phase, information and knowledge from previous studies were collected and possible data mining techniques were identified and used in the next phase.

Appropriate procedures were conducted to identify the best and the most suitable applicable technique(s). Testing on different techniques and methods were performed to see which one that works according to the requirements of the research elements.

The results of the study are then implemented in the last phase, observe and model organization. Two techniques of implementations were selected to prove that they would give great results for comparison. For mining part, a step-by-step approach was taken to accomplish the objective. Begin with fetching data from the source which is the social network website, then, data fetched is stored in a database. The data then was refined during the data preprocessing step before sending the data to be mined and analyzed using selected technique.

Finally, all of the information gathered through procedures and techniques implemented were analyzed, simulated and interpreted to discover the hidden pattern that can be discovered. This is also the phase where the result gained can be manipulated for other benefits.

Data Mining and Social Network Analysis Tools: Currently, there are many data mining and SNA tools that have been developed to implement analysis activities. Most of them are commercial products. However for the purpose of research, WEKA and AGNA are the most suitable tools. WEKA is a machine learning tool developed at the University of Waikato to design and develop data mining algorithms [18]. This tool enables computers to work based on collection of data (e.g. database). It is developed using Java and is a freeware under General Public License (GNU). In this research, WEKA was adopted in the implementation phase. Compared to other freeware tools, WEKA includes the whole range of data preparation. Moreover, its feature selections and algorithms are well integrated [19].

To conduct social network analysis, AGNA was used. AGNA is able to provide an intuitive framework that helps researchers to grasp the concepts and the philosophy underlies SNA besides its user friendliness [9].

Clustering: Clustering can be defined as a set or a group of data objects. It is characterized through a) similar objects in one cluster and b) different from other clusters [20]. Data retrieved are pre-processed and transform into comma separated value (csv) format to allow WEKA to cluster the processed data and produces groups of data which closest to each specific characteristics.

RESULTS AND DISCUSSION

First part of research, a set of questionnaire was distributed to a group of students in Universiti Teknologi Petronas, a chosen higher learning institution for this study. The respondents are from six different programs namely Chemical Engineering (CHE), Civil Engineering (CVE), Electrical Engineering (EE), Mechanical Engineering (ME), Petroleum Engineering (PE) and Computer and Information Sciences (CIS).

Through the survey, it is identified that Facebook has the highest percentage on most preferred SNS. The average time they spent on their favorite social network site is one to three hours per session. Based on the feedback, it is clear that socializing is the primary reason of joining social network services. This is followed by entertainment purposes, a medium of getting opinions, finding information, as well as sharing experiences and knowledge. However, it is very rare for them to use the services to get freebies or online gifts.

Table 1: The Analysis Result using COBWEB

	A	B	C	D
CHE	5	3	2	
CVE	6	4		
CIS	4	6		
EE		5		
ME	3	2	3	2
PE	3	2		

**A: Active, B: Moderate, C: Passive, D: Really Passive*

Table 2: The Clustering Result Using Simple Kmeans

Cluster	Respondent list
0 (10)	CVE10, CIS20, CIS19, CIS18, CIS17, CIS16, CIS15, CIS14, CIS12, CIS11
1 (13)	ME40, ME39, ME38, ME37, ME36, ME35, ME34, ME33, ME32, ME31, CVE7, CVE2, CIS13
2 (6)	PE50, PE49, PE48, PE47, PE46, CHE21
3 (8)	CVE8, CVE5, CHE30, CHE29, CHE28, CHE27, CHE25, CHE23
4 (10)	EE45, EE44, EE43, EE42, EE41, CVE9, CVE6, CVE4, CVE1, CHE26
5 (3)	CVE3, CHE24, CHE22

**In bracket: total number of instances in a cluster*

Other than that, an attractive discussion can attract thousands of users to participate in a mere minute thus increasing the chances of unearthing the potential for idea development from each comment posted. In higher education, an attractive discussion over academic subjects on Facebook may attract students to discuss their ideas. Therefore it is actually possible to classify students' characteristics with their consistency in the subjects discussed.

Mining the Social Network Data: In this second part of research, profiles of students from different backgrounds consist of education history, number of groups they joined, number of photos tagged, number of photo albums they have, links shared, notes created and program of study were collected manually and clustered. The objective of clustering is to analyze the relations between data collected and data processed. As a result, three clusters were created with each cluster member having similar characteristics.

There are two clustering analysis conducted; COBWEB (Hierarchical clustering) and Simple KMeans clustering. The analysis used a sample of 50 SNS users fetch from Facebook. The data collected were grouped into programs in the university.

COBWEB (Hierarchical Clustering): For each program, data that based on number of most shared features are extracted and stored in a database. Data extracted were number of notes written, links shared, photo albums created and status updated. These data are in numerical form and will be clustered based on most related to each other. The clusterer uses default acuity and cutoff values which both 1.0 and 0.002.

From the result in Table 1 below, out of 50 students, 42% are considered as active, while 44% are moderate. However, there are remaining few students who are clustered as passive and really passive. It could be because of their involvement in SNS is very minimal, thus making their distance measured too far from the other two groups.

The most common feature that is being shared by most active students is video links. Through the analysis and observation, students anticipate video as one of best mediums of studies. This is due to its practicality and interactivity compared to other mediums. Notes are the second item that most shared where it becomes an alternative way to share important articles or information. SNS provides interactive approach in collaborating feedbacks and responses. Another feature is messaging either through inbox or real-time discussion. This is faster than having a face-to-face meeting.

Simple KMeans clustering: KMeans clustering is a simple unsupervised algorithm to solve the common clustering problems [21]. By using simple KMeans clustering, dataset is clustered according to closest distance. From here, dataset that is most related to each other or almost similar will be grouped together without further clustering like hierarchical. This time, all dataset will be used for implementation.

The KMeans algorithm uses Manhattan distance as in order to find distance functions, 6 numbers of clusters was set and used the default seed value, 10.

Basically, every cluster is grouped by program. However, there are respondents from different programs are grouped under one cluster. As cluster 0, all CIS students are grouped together except student number 13. Compared to hierarchical clustering, all CIS student are common despite any range of usage. Through this second method of clustering, active, moderate or passive behaviors were not determined. The reason why KMeans clustering was selected is to compare the distance between all instances. Therefore, it reflects that CIS13 is an anomalous case because is closer to those in the ME group than those in the CIS.

As compared to the collected data (student profiles), obviously students from CIS (cluster 0) shared the highest number of notes while most CE students (cluster 5) shared the lowest number of notes. Moreover, by comparing clusters by links, it is analyzed that students in cluster 3 are the ones that share less links than any other students in other clusters. Therefore the number of sharing are varied despite the number of instances in a cluster remain constant. It can be concluded that student can fall under several categories and preferences based on their pattern in sharing behavior in SNS. This pattern however is clustered according to closest distances among each other.

Psychological and Personality Studies: There are several references can be implemented in finding characteristics and personalities of the users in the SNS. In this section, the discussion is focus on two foundations; a) the hierarchical clustering result that generates groups based on the users behavior. b) The simple KMeans clustering result that based on several behavioral principles affecting student's tendency on sharing online.

Cobweb: Hierarchical Clustering and Behavior: Based on the result, there are three types of SNS users in one cluster namely, active, moderate and passive. To support this result, Hans Eysenck's trait theories were referred. A trait means characteristics that an individual possess. When describing people through their behavior, it is equally as describing somebody's traits [22]. Based on Eysenck' theory, there are two main characteristics of people which are: a) extraversion and b) introversion.

Those who fall in the extraversion category are students who more open to society. In the network, they can be described as someone who loves to interact and be part of the network. In social network analysis, these students are the center or the bridge between other members in the network. Their active involvement in SNS is a reflection of their capabilities to be with anybody in their network. These students also reveal their traits through high number of sharing. For example, CVE4 and CIS14 who share many notes and links. An interactive mode of studies suits these students very well for example, real time online discussion.

Meanwhile, those who fall in the introversion category are students who have tend to be quieter and less open. They prefer to work on their own and social networking might be something that is not so important. Having an account on SNS might just be to follow the trend or playing computer games. Their characteristic reflected through number of activities they shared. As example, ME34 shared very minimal items on SNS.

Moderate students tend to be in either side such as CIS12 who had number of activities in between or balance. This type of students can be considered as bridge between the active and the passive in social network mapping. It is inaccurate to implement traits onto students merely through SNS account. A personal question or test should be given to these dataset in order to strengthen the result.

Kmeans Clustering and Behavior: In KMeans clustering, students are grouped based on distance. Whoever closes to each other, they are grouped together. It is hard to define who are the most active users or the moderates, or the passives. Therefore, a different theory is implemented.

Students are grouped according to how close they relate to each other. From this clustering result, each cluster is dominated by one program. Nevertheless, there are several students from outside programs join in. This can be seen as relation occurrence between programs in a university. From social network analysis point of view, this is a good finding to tell that there is a connection between students from different program from the pattern of SNS involvement.

Social Network Analysis Result: Based on Fig. 1, circled students are the centrals for each group. Degree centrality is based on overall distribution of Bavelas-Leavit (overall)

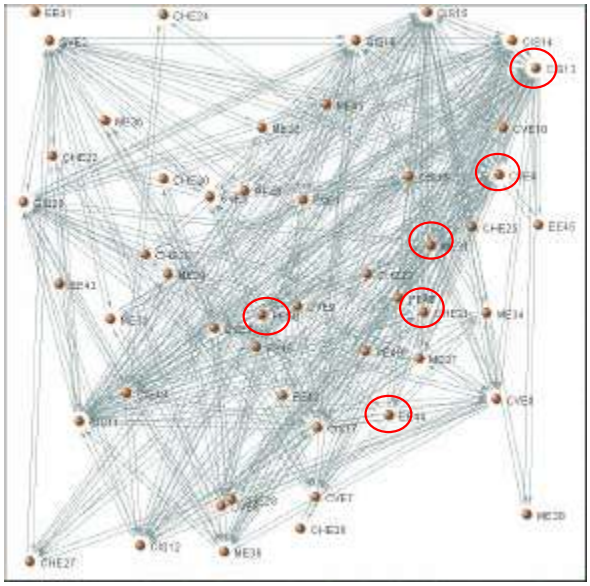

Fig. 1: Social Network Map

centrality coefficient. This finding is correlated with results gained from data clustering. The patterns of social network interactions can determine which student is playing a bigger role in collaboration with other students from other programs.

From here, potential interaction can be implemented for example, in Engineering Team Project; students are grouped from different programs to form a team to develop an engineering project. This analysis can be implemented in a way it describes which student is the best to be leader that has good interactions with other students who tends to work in silo or passive.

From the social map shown in Fig. 1, an analysis by group identifies CVE4, CIS13, CHE21, ME31, EE44 and PE50 are the center in each group. The results derived based on each instance interactions between other members in the group. Each of these instances shares almost similar characteristics (centrality value). Among the active users, they have many friends, befriending each other, using almost SNS features and updating SNS pages frequently. This result shall be interpreted more through a personality study that based on personality theory developed by Dr. Hans Eysenck.

Extroverts through SNS can be identified by several characteristics that the users may show. The following activities describe a student as an extrovert person; a) Happy to socialize, b) Participate extensively over the application – using almost all of the features, c) Sharing and discussing via wall post more than the passive (introverts), d) Get excited over responses through comments or posts, e) Best at wall/online discussion – quick response, f) Love to share videos, articles, notes and interesting posts and g) Online interactions provide a more dynamic social environment and more freedom to express themselves.

However, introverts tend to be otherwise. For example, they are quiet persons, tend to ignore major features in SNS, lack of updates and online collaborations is not the best medium of study for them.

From the analysis, all of the information gained from the social network map is compared based on Eysenck's personality theory. Through centrality analyzed, most of the centers are active users in SNS. Their characteristics display the standard of extroverts' characters. However, to strengthen the research, a real study is conducted over the samples using Eysenck's Personality Questionnaire (EPQ).

Eysenck's Personality Questionnaire: Most of the instances (students) display their true characteristics to almost as what they portray in SNS. For those earned high centralities in the SNA result, most of them portray extroversion. One of the most popular characters identified is the likeliness to have many friends and having interactions through social networking (e.g. wall chats). However, this result may not be accurate in all cases. This is due to some EPQ results are not aligned with the number of centralities gained in SNA. This reflects to some constraints of this research; a) Number of student interactions may be inaccurate since data entry is done manually, b) Traits are not easily portrays by the Internet, thus this research data is limited to whatever posted on the Internet only. Any real events may hardly being encountered, c) this research may not be applicable to those who have no SNS accounts, d) It is also requires consistent commitments from data members and e) The risk of social networking where information may be incorrect where some individuals may be an introvert by nature however, they get interactive through back end applications such as social network applications.

The Attractions of Social Network Sites: There are several factors that attract students to communicate and sharing knowledge online through SNS such as quick notification responses and self-expression opportunity. Therefore, forum and wall-to-wall discussion on SNS are two popular features that could be the effective ways for

online group studies. These features drive toward knowledge sharing culture among the students. They are motivated to share their photos, videos, notes and even their unprompted thoughts and opinions. It is suggested that SNS could be one of effective tools to cultivate creative group work and constructive feedback in higher learning environment.

CONCLUSION

This research does not aim to replace the traditional method of lectures at university. However it is more towards promoting dynamic study culture at the campus. Through this research, it has shown how data mining discovers the patterns of students' participation in social networking. It is found that their participation relates with their personal behavior. Further research would be the implementation on system where data matching can be manipulated to predict decision regards best knowledge practice.

REFERENCES

1. Boyd, D.M. and N.B. Ellison, 2008. Social Network Sites: Definition, History and Scholarship. Journal of Computer-Mediated Communication, 13(1): 210-230.
2. Pempek, A.T., A.Y. Yermolayeva and L.S. Calvert, 2009. College Students' Social Networking Experiences on Facebook. J. Appl. Development Psychol., 30(3): 227-238.
3. Nickson, C., 2009. The History of Social Networking. Retrieved February 17, 2010, from Digital Trends website: http://www.digitaltrends.com/features/the-history-of-social-networking.
4. Dwyer, C., S.R. Hiltz and K. Passerini, 2007. Trust and Privacy Concern within Social Networking Sites: A comparison of Facebook and MySpace. 13th Americas Conference on Information Systems, Colorado, USA, 9-12 August 2007.
5. Douglis, F., 2010. It's All About the (Social) Network. All System Go. Journal of IEEE Internet Computing, 14(1): 4-6.
6. Qiu, J., Z. Lin, C. Tang and S. Qiao, 2009. Discovering Organizational Structure in Dynamic Social Network. 9th IEEE International Conference on Data Mining, Miami, Florida, USA, 6-9 December 2009, pp: 932-937.
7. Weng, C.Y., W.T. Chu and J.L. Wu, 2009. RoleNet: Movie Analysis from the Perspective of Social Networks. Journal of IEEE Transactions on Multimedia, 11(2): 256-270.
8. Brandes, U., 2008. Social Network Analysis and Visualization. IEEE Signal Processing Magazine, 147.
9. Benta, M.I., 2005. Studying Communication Networks with AGNA. Journal of Cognitive, Creier, Comportament/Cognition, Brain, Behavior 2.1, 9(3): 574-567.
10. Van der Aalst, W.M.P. and M. Song, 2004. Mining Social Networks: Uncovering Interaction Patterns in Business Processes. Lecture Notes in Computer Science, 3080: 244-260. Springer. Retrieved from http://www.springerlink.com/index/lvw335vvkk5j1nyn. pdf.
11. Dekker, A., 2005. Conceptual Distance in Social Network Analysis. J. Social Structure, 6(3): 1-34.
12. Jensen, D. and J. Neville, 2002. Data Mining in Social Networks. National Academy of Sciences Symposium on Dynamic Social Network Modeling and Analysis, Washington D.C., USA, 7-9 November 2002, pp: 287-302.
13. Seifert, J.W., 2008. Data Mining and Homeland Security: An Overview. CRS Report for Congress, pp: 23-24.
14. DeRosa, M., 2004. Data Mining and Data Analysis for Counterterrorism. CSIS Report. The CSIS Press.
15. Luan, J., 2002. Data Mining and its Applications in Higher Education-Potential Applications. 42nd Annual Forum for the Association for Institutional Research, Toronto, Ontario, Canada, 2-5 June 2002, pp: 3-18.
16. Vialardi, C., J. Bravo, L. Shafti and A. Ortigosa, 2009. Recommendation in Higher Education Using Data Mining Techniques. 2nd International Conference of Educational Data Mining, Cordoba, Spain, 1-3 July 2009, pp: 190-199.
17. Cain, H.C., 0000. Investigating the effects of work context on the behaviors of medical care teams, with the goal of creating a computer-aided design tool for implementing clinical practice guidelines. Retrieved February 18, 2010, from Stanford Education website:http://www.stanford.edu/group/VDT/carol.htm.
18. Bouckaert, R.R., E. Frank, M.A. Hall, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, 2010. WEKA-experiences with a java open-source project. J. Machine Learning Res., 11: 2533-2541.

19. Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, 2009. The WEKA Data Mining Software: An Update; SIGKDD Explorations, 11(1): 10-18.
20. Williams, G., 2010. Data Mining Algorithms: Cluster Analysis. Retrieved August 19, 2010 from: http://www.slideshare.net/Tommy96/data-mining-algorithms.
21. Niknam, T., B.B. Firouzi and M. Nayeripour, 2008. An Efficient Hybrid Evolutionary Algorithm for Cluster Analysis. World Appl. Sci. J., 4(2): 300-307.
22. Kampen, V.D., 2009. Personality and Psychopathology: a Theory-Based Revision of Eysenck's PEN Model. Journal of Clinical Practice and Epidemiology in Mental Health, 5: 9-21.