Identification of User Patterns in Social Networks by Data Mining Techniques: Facebook Case

A. Selman Bozkır[1], S. Güzin Mazman[2], and Ebru Akçapınar Sezer[1]

[1] Hacettepe University, Department of Computer Engineering, Ankara, Turkey
selman@cs.hacettepe.edu.tr, ebru@hacettepe.edu.tr
[2] Hacettepe University, Department of Computer Education and Instructional Technologies, Ankara, Turkey
s.guzin@gmail.com

Abstract. Currently, social networks such as Facebook or Twitter are getting more and more popular due to the opportunities they offer. As of November 2009, Facebook was the most popular and well known social network throughout the world with over 316 million users. Among the countries, Turkey is in third place in terms of Facebook users and half of them are younger than 25 years old (students). Turkey has 14 million Facebook members. The success of Facebook and the rich opportunities offered by social media sites lead to the creation of new web based applications for social networks and open up new frontiers. Thus, discovering the usage patterns of social media sites might be useful in taking decisions about the design and implementation of those applications as well as educational tools. Therefore, in this study, the factors affecting "Facebook usage time" and "Facebook access frequency" are revealed via various predictive data mining techniques, based on a questionnaire applied on 570 Facebook users. At the same time, the associations of the students' opinions on the contribution of Facebook in an educational aspect are investigated by employing the association rules method.

Keywords: Social networks, decision trees, Facebook, association rules.

1 Introduction

In recent years, a rapid increase in numbers of social networks along with numbers of people using these networks has been observed. Social networks, also called social software or collaborative software, are a range of applications that augment group interactions and shared spaces for collaboration, social connections, and aggregate information exchanges in a web-based environment [1]. Similarly, [2] defined social networks as web-based services allowing individuals to 1) construct a public or semi-public profile within a bounded system, 2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system.

Millions of users have been interested in them since the introduction of social network sites (SNSs) such as MySpace, Facebook, Cyworld, Bebo, Twitter, etc. The majority of these users have integrated such sites into their daily lifes. Because most

S. Kurbanoğlu et al. (Eds.): IMCW 2010, CCIS 96, pp. 145–153, 2010.

of the social network users are young individuals, many of them are university students. Therefore, these sites are considered to play an active role in the younger generation's daily life [3], [4]. On the other hand, it has been stated that social networks have a prominent educational context, and this prominence has prompted a growing number of educators to consider them to be important sites for student learning although these are not intended primarily as educational applications. Besides, it has been suggested that these social networks help users re-situate learning in an open-ended social context by providing opportunities for moving beyond the mere access to content (learning about) to the social application of knowledge in a constant process of re-orientation (learning as becoming) [5].

There have been various studies about social networks in the educational context including using social networks as a tool or utilizing them as an environment for courses [6], [7], the utility of social networks in the teaching and learning process [8], their value for communication and collaboration [9], educational usage themes of social networks (e.g. [10], [11]). However, a study in the literature about data mining analysis of social network usage has not been encountered.

As one the most popular social networks, Facebook is considered in the present study. Facebook is defined as "a social utility that helps people share information and communicate more efficiently with their friends, family and co-workers" (facebook.com). As of November 2009, with 316 million users, Facebook is the most popular and well known social network throughout the world. Moreover, Turkey, with 14 million members, is the third country in terms of number of Facebook users and half of these members are younger than 25 years old [12].

Data mining is a process that uses a variety of data analysis tools to discover patterns and relations in data that may be used for prediction purposes. Supervised data mining techniques are used to model an output variable based on one or more input variables and these models can be used to predict or forecast future cases [13].

The purpose of the present study is to discover some usage patterns (i.e. usage time and access frequency) of Facebook users by data mining techniques. Additionally, an attempt is made to reveal the educational associations of the users. It is believed that social network based application development and educational programs can be enhanced by the findings of this study.

2 Data Mining

Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover useful patterns [13]. In other words, data mining is the complete process of revealing useful patterns and relationships in data by using techniques like artificial intelligence, machine learning and statistics via advanced data analysis tools. Oracle BI, SPSS Clementine, SAS Enterprise Miner and Microsoft Analysis Services are well known data mining tools in the marketplace [14].

Data mining methods are classified into two categories as predictive and descriptive. The aim of predictive methods is to make predictions on unseen cases by using

seen cases via a trained model. However, the goal of descriptive methods is discovering deep relationships, correlations and descriptive properties of data.

In this study, both of these method groups are employed by using SPSS Clementine 12. Additionally, various decision trees algorithms such as CART, CHAID and C5; artificial neural networks (ANN) and SVM (Support Vector Machine) classifiers in prediction of target variables are used. Furthermore, the variable importance feature of SPSS Clementine is used in discovering the factors affecting "Facebook usage" and "Facebook access frequency". Likewise, the Apriori algorithm is employed in discovering frequent opinions of students on the educational benefits of Facebook usage.

2.1 Methodology

As stated previously, various data mining techniques are employed during the analyses and except one (association rules mining discovery), their prediction performances are compared. Thus, in this section, a brief information is presented about the methodologies followed.

The decision tree method is probably the most popular classification method among the data mining techniques due to the ease of use and visual interpretation capabilities. Typically, a data mining task for a decision tree is classification; for example, to identify the credit risk for each customer [15]. The main idea of a decision tree is to split the data recursively into subsets so that each subset covers more or fewer homogeneous states of the dependent variable. At each split in the tree, all independent variables are recalculated for their impact on the dependent variable. When this recursive process is stopped and the tree is in a stable state, the required decision tree is formed [15]. At this stage, new cases can be classified via the decision tree. This stage is called tree deduction. C5, Quest, CHAID [16] and CART [17] are well-known decision tree algorithms. Nevertheless, SPSS Clementine serves whole algorithms in its package. In essence, differentiations among these algorithms are mainly caused by technical capabilities and employing different splitting approaches and their functions. For instance, C5 and CHAID algorithms are designed to classify only discrete valued variables by using "gain ratio" and "gini value" splitting approaches, respectively. However, CART algorithms are designed for both classification and regression purposes.

On the other hand, in the pattern recognition literature, SVM (Support Vector Machine) is a state-of-the-art method with its powerful discriminative features in linear and non-linear classifications. Generally, SVM is designed to enlarge the boundary of any two classes in pattern space by searching for an optimal hyper plane that has maximum distance to the closest points between two classes which are termed support vectors [18]. However, SVM has support for multiclass predictions via different developed kernel functions. By the help of these kernel functions, solving the problems in upper dimensional spaces becomes possible.

ANN are systems which contain intelligence nodes arranged in layers. In essence, an ANN has an input layer, a hidden layer, and an output layer. The nodes in the hidden layer collect the inputs from the input layer into a single output value which is

passed on to the output layer. Associated with each node in the network is a weight. The weights in the network are determined in a training phase of the network using training data. The network performance is then tested on the remaining data, or hold-out sample [19].

Association rule mining is again one of the best studied descriptive mining methods since the first design and creation. Agrawal, Imelinski and Swami stated a new approach to mining association rules in 1993 and designed a new algorithm, namely Apriori, via two phases seek mechanism on itemsets and by looking their association frequencies (Romero & Ventura, 2007). In the second stage of this study, the analyses are performed by using the algorithm Apriori. In association rules, mining analyzing, support, rule support, confidence and lift values are the important parameters in the usefulness evaluation of rules. In this study, lift and support values are considered.

Table 1. Variable names and available answers in the first part of the poll

Variable name	Type	Available answers and related distributions
Sex	Discrete	Male (50%) / Female (50%)
Age	Discrete	18-25 (74.1%) / 26-35 (20.53%) / 36-40 (3.86%) / 41 and above (1.4%)
Frequency of access to Facebook	Discrete	Once a year (0.18%) / Once a month (2.98%) / Several times a week (25.26%) / Once a day (22.81%) / Several times a day (48.77%)
Facebook usage time	Discrete	Less than 15 mins. (32.28%) / Half an hour (39.82%) / 1 hour (14.39%) / 1-3 hours (8.6%) / More than 3 hours (4.74%)
Education level	Discrete	High School (5.96%) / Bachelor (70.35%) / Master (23.16%)
Membership in any group	Discrete	Yes (99.82%) / No (0.18%)
Membership in student groups	Discrete	Yes (86.49%) / No (13.51%)
Membership in common interest groups	Discrete	Yes (77.54.5) / No (22.46%)
Membership in internet & tech groups	Discrete	Yes (27.02%) / No (72.98%)
Membership in organizations	Discrete	Yes (61.93%) / No (38.07%)

3 Data

Data was collected from 570 active Turkish Facebook users (students) with an online poll. This online poll consisted of two sections. In the first section, demographic characteristics of Facebook users and their frequency of Facebook usage, length of time spent on Facebook, and memberships in Facebook groups were collected. In the second section, a 10-point Likert scale with 11 opinions were asked, the answers ranging from 1 (strongly disagree) to 10 (strongly agree), like “Facebook contributes to communication between classmates”, “It’s useful for assigning tasks in classes and

homework assignments". Thus members' views of Facebook in relation to its educational usage were sought.

The variable names of the first part and available answers are given in Table 1. Although the initial dataset size was larger than 570 people, during the data cleaning and transforming steps, 13 people were removed due to the absence of sufficient information. Therefore, the final dataset comprised 570 people. In the dataset, male and female participants are almost equal and more than 400 applicants are in the 18-25 age range. Furthermore, almost all students are at either undergraduate or graduate level.

4 Application of Data Mining

To discover important factors that affect Facebook usage time and access frequency to Facebook, CART, CHAID, C5, artificial neural network and SVM algorithms, which are built in to SPSS Clementine 12, were employed on the dataset at hand (see Fig. 1). The overall data is partitioned as 80% training and 20% testing, respectively. Training and test datasets are selected randomly. As the dataset consists of discrete valued variables, the true and false prediction rates are listed.

According to the results (see Table 2), SVM achieves the most accurate predictions for two target variables. Therefore, it is considered that the variable importance results of SVM are the most accurate predictions. As can be seen in Fig. 2, sex, education level, membership in a group and membership in any common interest groups are the most important factors affecting *Facebook usage time*. Sex plays a crucial role in Facebook usage time with 68%. Again, it can be clearly seen that age, membership in student groups and usage time variables are the most important factors affecting *access frequency to Facebook*. The effect of age is more than 80% in *access frequency*.

Table 2. Applied algorithms and prediction results

Target variable - Applied algorithm	True classification	False classification
Facebook usage – SVM	62.63 %	37.37 %
Facebook usage – ANN	47.72 %	52.28 %
Facebook usage – C5	47.54 %	52.46 %
Facebook usage – CART	43.68 %	56.32 %
Facebook usage – CHAID	41.40 %	58.60 %
Access frequency to Facebook – SVM	69.65 %	30.35 %
Access frequency to Facebook – C5	55.79 %	44.21 %
Access frequency to Facebook – CART	52.81 %	47.19 %
Access frequency to Facebook – CHAID	50.35 %	49.65 %
Access frequency to Facebook – ANN	48.77 %	51.23 %

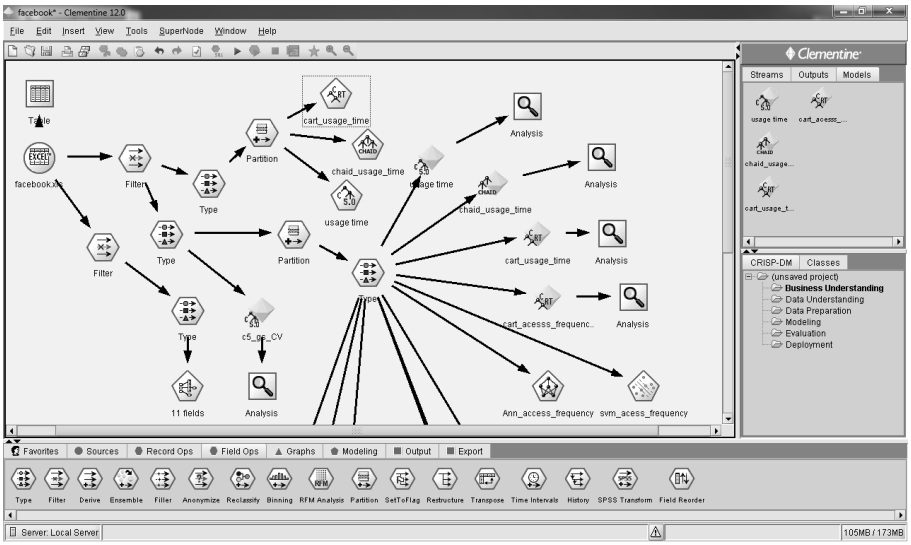

Fig. 1. Applying data mining methods in Clementine 12

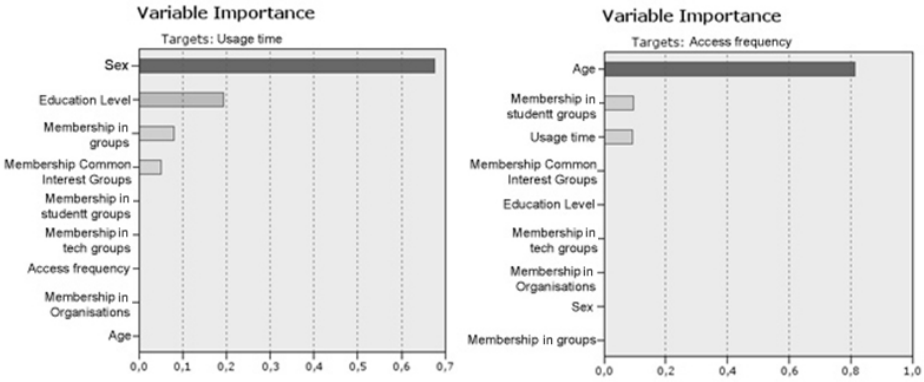

Fig. 2. Variable importance values of two target variables

On the other hand, in the association rules mining study, the association of the student ideas on Facebook and its educational benefits has been investigated. To achieve this, the well-known Apriori algorithm was run with 5% support and 15% confidence parameters.

As can be seen in Table 3, some interesting rules are listed sorted by lift values. Lift value shows the usefulness and attractiveness of a rule. Therefore, the rules which have lift values higher than 1 should be considered carefully for educational purposes.

Table 3. A sample subset of discovered association rules

Antecedent	Consequent	Confidence	Support	Lift
"It contributes to the communication between teacher & student" = 7	"It's useful at accessing the rich learning resources" = 9	11.32%	1.03%	3.0
"Facebook contributes to communication between classmates" = 4	"It's useful for executing the group tasks" = 2	12.83%	0.87%	2.9
"Facebook contributes to communication between classmates" = 8 and "It contributes to communication between teacher & student" = 8	"It contributes to transferring course materials and resources" = 6	20.69%	1. 05%	2.7
"Facebook contributes to communication between classmates." = 6	"It's useful at providing rich multimedia contents in teaching"= 3	20.51%	1.40%	2.65
"It contributes to the communication between teacher & student" = 3	"Facebook contributes to communication between classmates" = 6	17.02%	1.40%	2.48
"Facebook contributes to communication between classmates"=7 and "It contributes to communication between teacher & student"=5	"It contributes to dissemination of announcements of lectures & classes" = 2	12.5%	0.52%	2.03

5 Discussion and Conclusion

This study tried to discover the factors affecting access frequency and usage time of Facebook by various decision tree algorithms, ANN and state-of-the-art algorithm SVM. According to the results, SVM exhibits the most accurate results due to the nature of the dataset at hand. It is believed that the prediction capabilities can be enhanced by using more training data. On the other hand, the associations of the student ideas were explored by employing the Apriori algorithm and, as can be seen from the results obtained, the contribution of Facebook to communication between classmates is more than to communication between students and teachers. Moreover, the students who hold these views believe that Facebook is a good medium for accessing rich resources. More of these types of rules can be revealed by using the Apriori algorithm and the use of social network sites for educational ends can be reformed in the light of these rules.

If the increasing trend in social network sites usage is considered, the importance of applications and approaches related to social networks can be easily understood. Targeting specific ages or sex may strategically affect the success of developed applications. As a consequence, data mining methods can be successfully employed on social network usage data.

References

1. Bartlett-Bragg, A.: Reflections on Pedagogy: Reframing Practice to Foster Informal Learning with Social Software (2006), http://www.dream.sdu.dk/uploads/files/Anne%20Bartlett-Bragg.pdf
2. boyd, D.M., Ellison, N.B.: Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication 13, 210–230 (2007)
3. Lenhart, M.: Adults and Social Network Websites. Pew Internet & American Life Project Report (2009), http://www.pewinternet.org/pdfs/PIP_Adult_social_networking_data_memo_FINAL.pdf
4. Bumgarner, B.A.: You Have Been Poked: Exploring the Uses and Gratifications of Facebook Among Emerging Adults. First Monday, 22 (2007), http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2026/1897
5. Mejias, U.: Nomad's Guide to Learning and Social Software (2005), http://knowledgetree.flexiblelearning.net.au/edition07/download/la_mejias.pdf
6. English, R., Duncan-Howell, J.: Facebook Goes to College: Using Social Networking Tools to Support Students Undertaking Teaching Practicum. Journal of Online Learning and Teaching 4, 596–601 (2008)
7. Lockyer, L., Patterson, J.: Integrating Social Networking Technologies in Education: A Case Study of a Formal Learning Environment. In: Proceedings of 8th IEEE International Conference on Advanced Learning Technologies, Spain, pp. 529–533 (2008)
8. Ajjan, H., Hartshorne, R.: Investigating Faculty Decisions to Adopt Web 2.0 Technologies: Theory and Empirical Tests. The Internet and Higher Education 11, 71–80 (2008)
9. Saunders, S.: The Role of Social Networking Sites in Teacher Education Programs: A Qualitative Exploration. In: McFerrin, K., et al. (eds.) Proceedings of Society for Information Technology and Teacher Education International Conference, pp. 2223–2228. AACE, Chesapeake (2008)
10. Mazman, S.G., Usluel, Y.K.: Adoption Process of Social Network and Their Usage in Educational Context. Master Thesis. The Institute for Graduate Studies in Science and Engineering. Hacettepe University, Ankara (2009)
11. Selwyn, N.: Web 2.0 Applications as Alternative Environments for Informal Learning - A Critical Review. Alternative Learning Environments in Practice: Using ICT to Change Impact and Outcomes, OECD-KERIS Expert Meeting (2007)
12. Check Facebook (2009), http://www.checkfacebook.com
13. Berry, M., Linoff, G.: Mastering Data Mining: The Art and Science of Customer Relationship Management. John Wiley & Sons, Chichester (2000)
14. Bozkir, A.S., Gök, B., Sezer, E.: İnternetin Eğitimsel Amaçlar için Kullanımını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti. In: Bilimde Modern Yöntemler Sempozyumu, pp. 833–842. Eskişehir (2008)
15. Tang, Z., MacLennan, J.: Data Mining with SQL Server 2005. John Wiley & Sons, Indiana (2005)
16. Kass, G.V.: An Exploratory Technique for Investigating Large Quantities of Categorical Data. Applied Statistics 29, 119–127 (1980)
17. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, P.J.: Classification and Regression Trees. Wadsworth International Group, Belmont (1984)

18. Liu, J., Wang, Z., Xiao, X.: A Hybrid SVM/DDBHMM Decision Fusion Modeling for Robust Continuous Digital Speech Recognition. Pattern Recognition Letters 28, 912–920 (2007)
19. Fuller, C.M., Piros, D.P., Wilson, R.L.: Decision Support for Determining Veracity via Linguistic-Based Cues. Decision Support Systems 46, 697–703 (2009)
20. Romero, C., Ventura, S.: Educational Data Mining: A survey from 1995 to 2005. Expert Systems with Applications 33, 135–146 (2007)