Educational Social Network Group Profiling: An Analysis of Differentiation-Based Methods

João Emanoel Ambrósio Gomes[1], **Ricardo Bastos Cavalcante Prudêncio**[1]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Caixa Postal 7851 – 50.732-970 – Recife – PE – Brazil

{jeag,rbcp}@cin.ufpe.br

Abstract. *Social media communities are usually formed by similarities among users. In educational social networks, several factors propitiate the user group generation, e.g. share the same academic environment or interested in common curricular. In order to explain the group formation resulted from educational social network, we applied two group profiling methods based on differentiation. Wilcoxon rank-sum test and PART rules algorithm were applied to a dataset available, the OJE educational social network. The performed experiments showed that the methods were effective to group profiling generation, characterizing 81.81% and 100% of groups, respectively.*

1. Introduction

Great advances have been observed in the educational systems in recent decades as a result of the adoption of various technologies such as online learning platforms, intelligent tutoring systems, educational games and educational social networks (ESNs) [Ha et al. 2000]. Organizations that adopt such systems are able to collect large amounts of data on web servers, databases and access logs, in various formats. Data repositories contain information that can be useful for modeling and evaluation of the learning process as well as aiding decision making by managers and administrators of educational systems [Romero and Ventura 2007][Mostow and Beck 2006].

Some techniques of data mining are especially useful in the scope of electronic educational systems [Romero and Ventura 2007]. In this context, the educational data mining rises as an emerging field that involves application of computational techniques to identify patterns in large educational data repositories [Baradwaj and Pal 2012] [Farzan and Brusilovsky 2006].

Data mining currently can handles the problem of explore structures with lots of information (properties) and heterogeneous datasets (e.g. social networks). Several works reported statistical patterns which were presented in complex networks across many domains [Newman 2003] [Sun et al. 2007]. On the other hand, this present work focuses on analysis of groups (communities) in social media. In particular, we build descriptive profiles of student groups in an ESN, intending to explain its formation.

Common interests or affinities encourage the formation of communities in educational environments [Baradwaj and Pal 2012]. For instance, some users may interact because they share the same school/classroom, are engaged in the same activities or are interested in the same study subject or course. Identifying the features that distinguish a group to others in the network is important to explain the dynamics of group formation and also support the decision making process in the education environment.

In this context, this paper proposes two differentiation-based methods to perform group profiling in ESNs, which can help to understand the community formation process on online educational platforms. Our goal is to extract attributes of each individual user and verify whether the group members really have interests and/or common characteristics that differentiate them from the rest of the network.

In this strategy, we initially applied the Multi-level Aggregation Method [Blondel et al. 2008] for discovering groups in the social network data. Giving a set of attributes that describes the users, the groups profiles are produced by using the Wilcoxon rank-sum test and supervised machine learning techniques. The first method identifies those attributes which presents a distribution within the group that is statistically different from the distribution observed in the rest of the network, identifying the best discriminative attributes. The Wilcoxon test establishes a difference between two sample using magnitude-based ranks.

Although the effectiveness of group profiles generation using the Wilcoxon test shown in [Gomes et al. 2013], it was not able to characterize groups that showed no statistically significant differences from the rest of the network. Aiming to improve the coverage of groups in the profiling process, we propose the application of supervised machine learning techniques. Here, the part rules algorithm was applied to generate the profiles from the set of selected rules.

Experiments were performed using data collected from the ESN OJE[1]. The OJE is a web platform that works as a social network, where users are presented to challenges in the form of games and questions about several school subjects, called enigmas [Meira et al. 2009]. Based on the logs generated by the activities (games and enigmas) and users' personal information, it is possible to collect attributes that enable the group profiling study of this social network.

The remaining of this paper is organized as follows: In Section 2 we describe the group profiling strategy, followed by Section 3, where we introduce the experimental settings. In Section 4, the experiments performed and a discuss of the results obtained are presented. In Section 5 we present some related work. Finally, in Section 6 we conclude and point some future works.

2. Profiling Strategies

We apply the same group profiling strategy adopted in [Gomes et al. 2013], adding a group profiling method (PART). Figure 1 presents the general process followed by the strategy for group profiling. First, the data set are preprocessed, for extracting features to **User's Representation**. After that, the network structure is produced, composed by a set of nodes (representing the users) and their corresponding edges, for that we used the tool Gephi[2]. A community detection method is applied on network for identifying the existing communities (or groups). In this step, we adopted the Multi-Level Aggregation Method algorithm (MAM) [Blondel et al. 2008] for **Communities Detection**, since the groups are not explicit in the network. Finally, the **Group Profiling Method** is applied to identify relevant features that discriminate each group. In this work, we applied

[1] http://www.acre.oje.inf.br/oje/app/index

[2] https://gephi.github.io/

two differentiation-based group profiling methods: the Wilcoxon rank-sum test and PART rules algorithm.

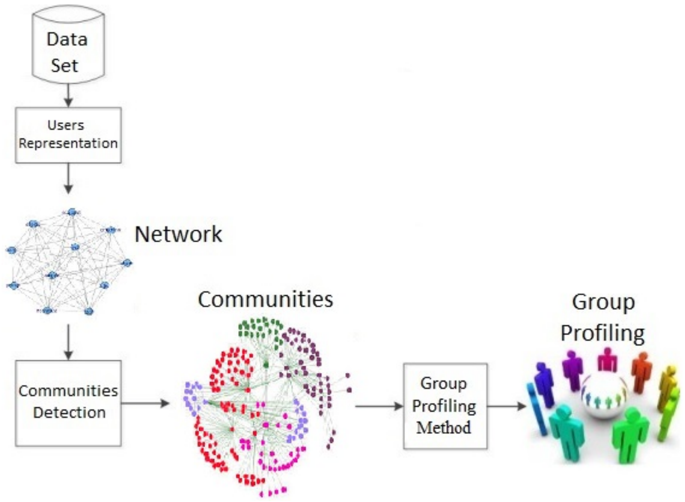


Figure 1. An overview of the group profiling strategy [Gomes et al. 2013].

The selected method, PART, produces as outputs a set of rules easily understandable by humans, facilitating the understanding of the groups profiling. The PART algorithm is a variation of the J48, implementation of the algorithm C4.5 which generates a decision tree, which builds production rules from decision tree. The process of generating production rules operates in two steps, rules are initially induced as tree and are further refined. For each rule created is estimated the coverage of instances. This occurs repeatedly until all instances are covered. The rules with higher coverage are presented to the user and the others are discarded [Hall et al. 2009].

For this work, we used the learning algorithms to solve a binary classification task for each group. The class label indicates if a user belongs or not to a given group. The PART algorithm will learn rules that separate the users in a group from the rest of the network. In the next section, we present the case study and experimental methodology adopted to evaluate our proposed approach.

3. Case Study and Experiment Setup

In this section, we present the evaluation of our strategy to group profiling applied to an ESN called OJE.

3.1. OJE

The OJE is a social network that connects students and teachers through games and enigmas. As main activities the OJE includes: (i) games with known mechanical youth to ensure their motivation, (ii) constructed enigmas in the format of the Brazilian National High School Exam's questions.

The OJE's platform provides an environment for conducting tournaments between teams and individual students (supervised by teachers) who engage in various disputes. The project also enhances the teaching processes through areas dedicated to teachers, such as a section of lessons tips to help them use the games in their disciplines, and a bank of questions (in development) that facilitates the composition of exercises from the enigmas.

3.2. Data Set

As aforementioned, to conduct a group profiling study, a suite of related data on individual attributes is necessary. Hence, we selected the OJE social network data in our case study. The OJE network presents 5590 users, of which 5204 are active with 9340 relationships. In Table 1, we detail the network information.

Table 1. Statistics on OJE

#Active Users	5204
#Links	9340
Link Density	0.001
Average Link	3.59
Diameter	12

3.3. Users' Representation

For users' representation, we performed several pre-processing procedures. Initially, a set of 40 individual attributes was extracted from the database. It was selected a set of 13 educationally descriptive attributes. The attributes School, Grade and City, were removed as these imply in obvious groups. The selected features are described below.

- Age: This attribute was used to verify the existence of groups by age ranges.
- Access: We applied a verification of users activity level by establishing three attributes: Website, games and enigmas. These were extracted from server logs.
- Participation in enigmas: Aiming to analyze the participation on enigmas, three attributes were defined: the questions accessed number, and number of correctly and incorrectly answered enigmas.
- The Classification of games and enigmas by related educational area: In OJE, each game and enigma has a classification that defines its educational area. There are six attributes used to group game and enigmas access number. Both games' and enigmas accesses were distinguished by Nature, Literature and Humanities.

3.4. Community Detection

None explicit community has been defined yet in OJE social network. Thus, it demanded the application of external algorithms to identify communities groups. We started the pre-processing of the data by removing the singletons (single node) from the dataset. Since the objective is to build group profiling, they could not be in any community. Using the Multi-level Aggregation Method [Blondel et al. 2008], we identified 29 groups.

Groups that had fewer than 10 users were removed, since they were considered too small and irrelevant for the study. We calculated the density for each group, as it is a common metric of how well connected a network is (i.e., how closely knit it is) [Tang et al. 2011]. Only 10 groups were selected, based on their density values. As the 10th and 11st showed the same density value, we added the last. The Table 2 shows the preprocessed database statistics, in which is presented the number of users and links, the density, the network average degree, the network diameter and the groups number.

Table 2. Statistics on OJE Pre-processed

#Active Users	227
#Links	672
Link Density	0.026
Average Link	5.921
Diameter	8
Group Numbers	11

In Figure 2, we visualized the resulting network after the pre-processing step. Analyzing the figure we can separate groups formed in the network, and identify their labels. In Table 3, we have all the statistics of each group individually, introducing size, imbalance ratio[3], density, and average degree of each group. Table 3 shows that the imbalance ratio is significantly different for each group and we can identify communities that are more cohesive than others, for example, comparing the group 25 and 19.

Table 3. Statistics on Groups

Group	Size	Imbalance Ratio	Average Degree	Density
1	12	17.9	2.5	22.7%
2	23	8.9	3.217	14.6%
3	26	7.7	3.692	14.6%
4	13	16.5	3.538	29.5%
12	19	10.9	3.368	18.7%
15	23	8.9	4.0	18.2%
17	24	8.5	4.75	20.7%
19	14	15.2	2.286	17.6%
20	20	10.4	2.8	14.7%
25	28	7.1	5.857	21.7%
28	25	8.1	5.84	24.3%

3.5. Group Profiling

After the communities detection, two differentiation-based group profiling methods were applied: Wilcoxon Rank-sum test and the PART rules algorithm. In differentiation-based methods we selected features which differs one group from others in the network.

We adopted the same configuration used in [Gomes et al. 2013] for the Wilcoxon test, therefore, the method works by pairing the distribution of the attribute values of a particular group comparing to the values of the remaining groups. As the goal of this method is to analyse whether there are differences among the groups in relation to rest of the network, for each pair of attribute (group vs rest of the network), the hypothesis is tested: the median of the attributes are equals or different? If the null hypothesis (H_0) is rejected (p-value $\leq \alpha$), the attribute has statistically significant difference and is selected for the profile the group, otherwise the attribute is not selected. It was considered a

[3] Is the ratio between the size of the rest of the network (majority/common) and group (minority/rarest).

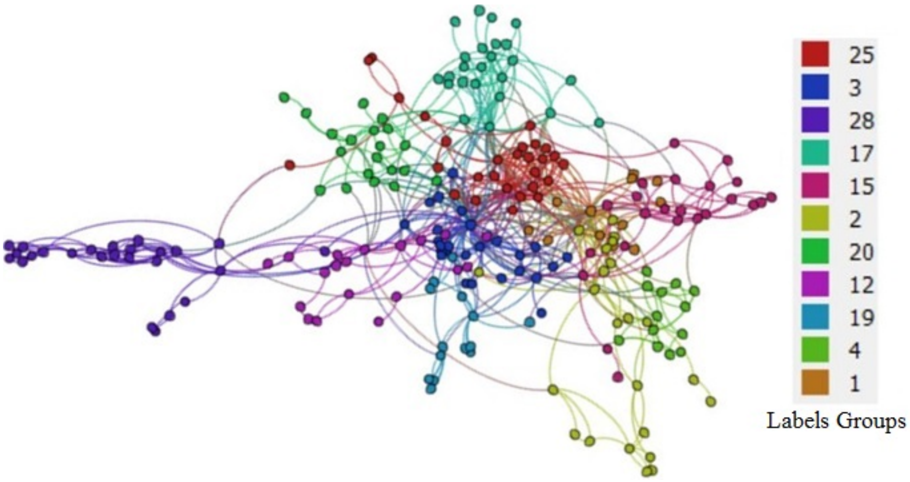


Figure 2. Network resulting from pre-processing.

significance level of significance level 5% (α = 0.05). The group profile is the list of features that characterize the community according to the statistical test. The Figure 3 illustrated the decision process carried out to verify the hypothesis test.

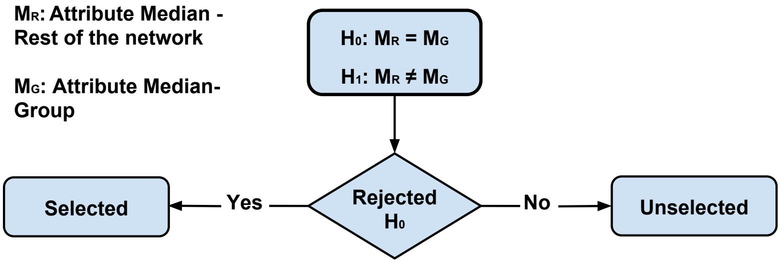


Figure 3. Decision-making process of the Wilcoxon rank-sum test.

For the supervised learning method the group profiling problem amounts to rules selection in a 2-class classification problem with the group being the positive class and the remaining nodes in the network as the negative class. The goal is to select a set of rules generated by the PART algorithm that describes each community. Figure 4 shows the process of generating group profiling of the supervised learning method.

As indicated in Table 3 the imbalance ratio is significantly different for each group (sample), as solution we apply undersampling to produce a random subsample. This filter allows to specify the maximum ”spread” between the rarest and most common class. All experiments of this work were executed using the Waikato Environment Knowledge Analysis framework (WEKA). We used the implementation of undersampling, Spread Subsample of the WEKA. We defined the maximum class distribution spread (M) parameter of Spreead Subsampleas as 1. This way, we obtained a uniform distribuition (1:1), i.e., the algorithm randomly selects the same number of negative (rest of the network) instances from the positive set (group). For the experiments with PART, we used the the WEKA default parameters.

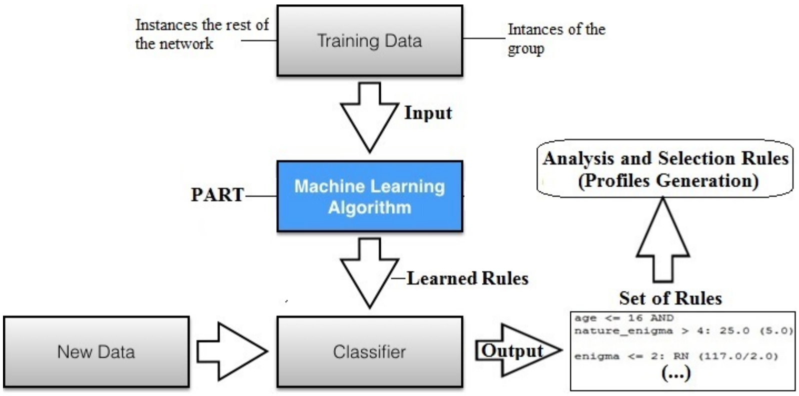

Instances the rest of the network
Training Data
Intances of the group
Input
PART
Machine Learning Algorithm
Analysis and Selection Rules (Profiles Generation)
Learned Rules
Set of Rules
New Data
Classifier
Output
age <= 16 AND
nature_enigma > 4: 25.0 (5.0)
enigma <= 2: RN (117.0/2.0)
(...)

Figure 4. Process generating group profiling supervised learning.

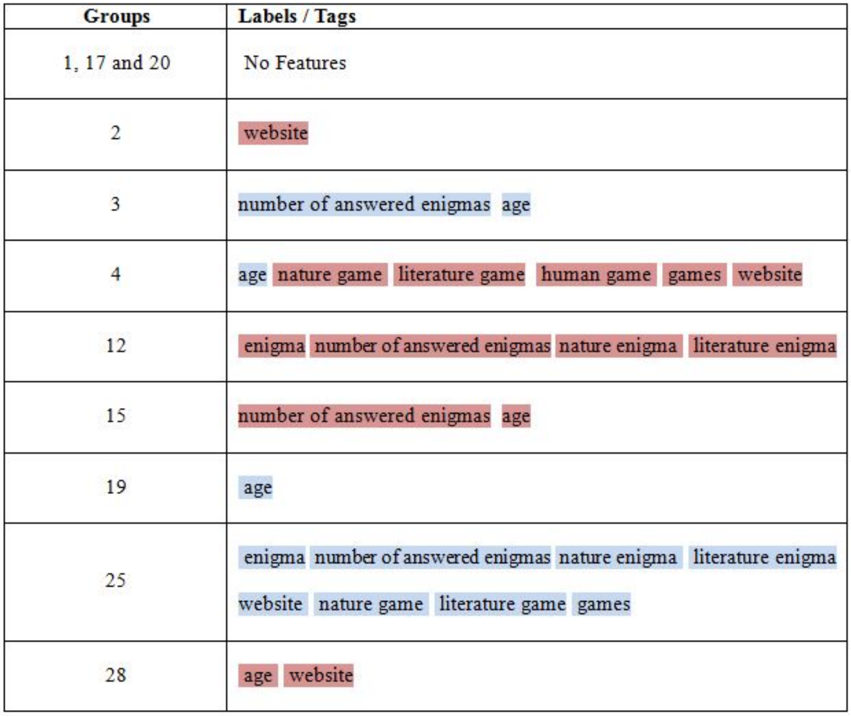
4. Results and Discussion

In this section, we present the results obtained by the two group profiling methods, applied to characterize the users identified groups on OJE network by the Multi-level Aggregation Method [Blondel et al. 2008].

4.1. Wilcoxon Rank-sum Test

In Table 4, we can visualize the relevant features for each group, according to the Wilcoxon test results.

Table 4. Profiles of Each Group - Wilcoxon Test



Groups	Labels / Tags
1, 17 and 20	No Features
2	website
3	number of answered enigmas age
4	age nature game literature game human game games website
12	enigma number of answered enigmas nature enigma literature enigma
15	number of answered enigmas age
19	age
25	enigma number of answered enigmas nature enigma literature enigma website nature game literature game games
28	age website

We selected only the features that presented statistically significant differences according to the Wilcoxon test. The features are marked in blue when its average value in the group is greater than the average feature value observed in the rest of the network.

The red marked in turn indicates that the features within the group has a lower average compared to the average value considering all network users.

The combination of the identified features for each group is unique, which demonstrates the potential of the proposed method to characterize the groups. In order get a better understanding of the results, here we discussed three concrete examples: the groups 12, 20 and 25.

As shown in Table 3, we found that the group 25 is the most prominent one in the network. Figure 5 (a) presents a bar plot containing the average value of each attribute describing group 25 compared to the average value for the rest of the network. We observed that the group 25 consists of people who are very involved in OJE, from website usage to the participation in games and enigmas.

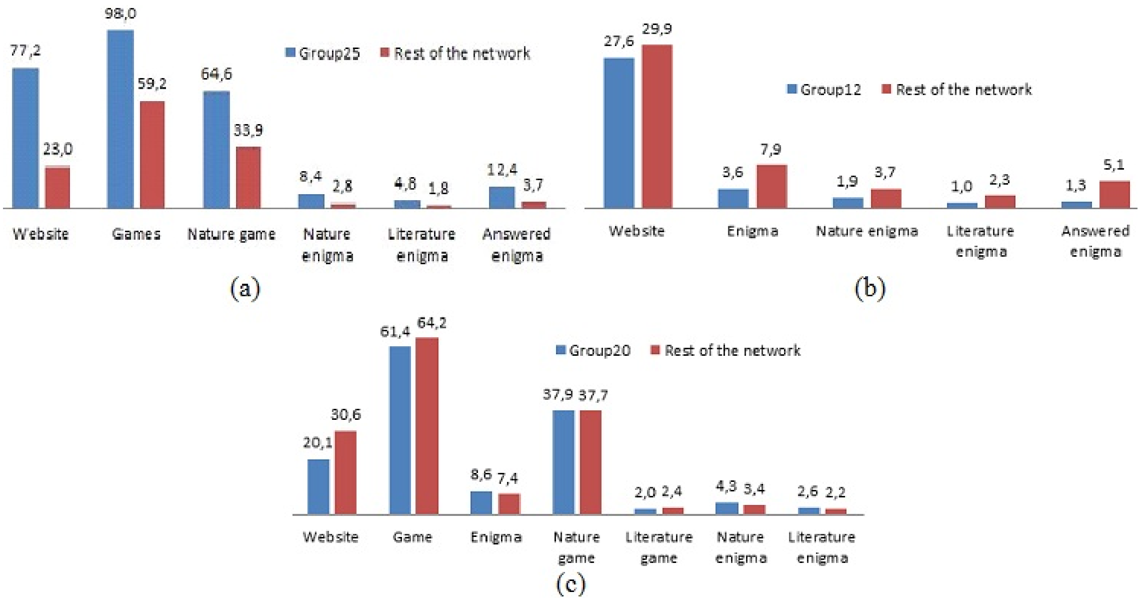


Figure 5. Averages of attribute values of one group (blue) and the rest of the network (red). (a) we see that the group 25 really stands out about the rest of the network (b) Despite of group 12 have a good OJE access average, have little access to enigmas (c) Observe that there are no significant differences attributes between Group 20 and rest of the network.

Regarding group 12, we can see the opposite behavior, as it can be seen in Figure 5 (b). All features describing group 12 present an average value that is lower than the observed average for the rest of the network. This group of users is a tied community, not only by the structure identified by the community detection method, but also by common behavior. This group has a good average (although not significant) access to the website, but very low level of access to enigmas, which is not interesting for the educational purposes of OJE.

The limitation of the proposed group profiling approach with the Wilcoxon test is in those groups where the differences between the attributes of groups and the rest of the network are minimal (not significant). In such cases, the test did not reveal any feature to categorize the group, showing not be effective in such cases. In Figure 5 (c), we can see an example of the group 20.

4.2. Supervised Machine Learning

With the aim of improve the groups coverage in the profiling process, we propose the application of PART rules algorithm. For ensure that every subsamples from the original dataset has the same chance of appearing in the training and testing set, we apply corss-validation with 10 folds (default parameter WEKA). The Table 5 shows the accuracy achieved by the PART algorithm in each experiment performed with the subsamples of each group. The overall average accuracy of the PART algorithm for the eleven experiments was 58.62%.

Table 5. Accuracy Achieved by the PART Algorithm in Each Experiment

Group	1	2	3	4	12	15	17	19	20	25	28
Accuracy(%)	58.3	63	53.8	69.2	39.5	58.7	56.25	50	62.5	69.6	64

The next step in our strategy was to extract interpretable knowledge from the rules generated by the classification algorithm. Our goal is to discover rules to support the communities formation understanding. The PART algorithm generated rules for all groups, overcoming the Wilcoxon test (81.81%).

A total of 58 rules were generated, but only rules with coverage $\geq$ 10% of the subsample (network) users were selected for analysis. The number of rules generated for each group and the selected rules for analysis, can be seen in Table 6. They are presented in a pseudo-code structure to aid the understanding. Observing the generated rules, we can check the suitability of classification algorithm for group profiling, as well as the actual relations among the profiles generated by the Wilcoxon test and the selected rules.

In Table 6, each line shows the group indicated (1, 2, 3, 4, 12, 15, 17, 19, 20, 25 and 28), the number of generated rules, the total of instances classified by each rule and the misclassified (in parentheses). Observing the Table 6, for group 1, the PART algorithm generated one rule that classified 8 instances without error, i.e., 100 % of accuracy. Whereas for group 15, the single rule generated identified 13 instances, but two did not belong to group (84,61% of accuracy).

Observing the accuracies of the rules, we found that 9 of the 15 selected rules obtained 100% of accuracy in classification. Highlighting the groups not characterized by the Wilcoxon test, we found that the PART algorithm was effective in generating profiles of these groups. As described in the rules, the group 1 is a community with good access to the website and low access on enigmas; the group 17 has a low access to enigmas and games, however well distributed; Finally, the group 20 consists of users between 16 and 17 years with good access to the website and games.

Other identified features with the Wilcoxon test were also observed in the PART algorithm such as: the high age in groups 3 and 15, low access to literature enigmas by group 12; considerable access to enigmas by group 25; and finally the lower age group 28.

Observing the results in general, we check the low access to enigmas by the general network, showing that access to OJE are targeted more for games and website. Analyzing the profiles generated by the two methods, we realize that some groups that not well described by the Wilcoxon test are best detailed by PART (groups 1, 17, 20 and 19), and unlike for the groups 25 and 12. For the group 2, the Wilcoxon test identified a good access to the website and the PART algorithm detailed this presence with access to human games and enigmas.

Table 6. Analysis and Selection Rules PART Algortihm

Group	Number of Rules	Selected Rules
1	1	website >20 AND correctly_enigmas <= 3: **1 (8)**
2	4	age <= 17 AND correctly_enigmas <= 0 AND human_game >1: **2 (10/1)** human_enigma >1: **2 (7)**
3	3	age >14 AND literature_game <= 8 AND enigma >4 AND human_enigma <= 3: **3 (9)**
4	1	incorrectly_enigmas <= 0 AND literature_enigma <= 0 AND website >4: **4 (5)**
12	1	literature_enigma <= 1: **12 (28/10)**
15	3	age <= 18 AND games <= 90: **15 (13/2)**
17	2	literature_game <= 7 AND human_enigma <= 1: **17 (9)** literature_enigma <= 4 AND nature_enigma <= 2 AND correctly_enigmas <= 1 AND human_enigma <= 1 AND answered_enigmas <= 0 AND games <= 2: **17 (5)**
19	1	literature_enigma <= 2 AND human_enigma <= 1 AND nature_enigma <= 2: **19 (13/3)**
20	3	age >15 AND age <= 16: **20 (10/1)** age <= 17 AND website <= 28 AND games >15: **20 (4)**
25	1	enigma >2: **25 (33/7)**
28	3	age <= 17 AND literature_enigma >1: **28 (10)** age <= 17 AND nature_enigma <= 1: **28 (8)**

5. Related Work

The Group profiling describes shared characteristics of a group of people. According to [Tang et al. 2011], the main group profile objective is to understand the formation of explicit or implied communities, using individual attributes. In this work, three sensible methods of group profiling are presented in a comparative study: aggregation, differentiation, and egocentric differentiation. This work uses individual attributes for group profile. In [Gomes et al. 2013],the authors adopt the Wilcoxon rank sum test [Mei et al. 2008] as a differentiation-based group profiling method. In our study, we analyse the effectiveness the application of supervised machine learning in group profiling, seeking to improve to coverage of groups in the profiles process.

Another line of research relevant to group profiling is to extract annotations from relational data with text. [Chang et al. 2009] proposes the NUBBI (Networks Uncovered By Bayesian Inference) to infer descriptions of entities in a text corpora as well as relationships between these entities. The probabilistic topic model assumes the words are generated based on the topics associated with an entity or the topics of the pairwise relationship of entities. NUBBI annotates connections, rather than groups as we do in this work.

6. Conclusions and Future Work

Several methods are available for communities detection in an ESN. We adopt two group profiling methods to find out possible reasons that causes formation of a community. This insights help to explain why the users connect and interact among them in ESN.

In this study we present two differentiation-based group profiling methods: Wilcoxon rank-sum test and PART rules Algorithm. Despite not indicating any descriptive feature in case of groups with not statistically significant differences, the Wilcoxon test was effective in identifying tags to characterize the groups. In fact, descriptive features were identified for 81.81% of the groups. As seen in the result analysis, the labels identified by the test became good profiles for groups.

Due to the imbalance between the instances of the groups relative to the rest of the network, data were preprocessed with undersampling (M=1). The PART algorithm characterized all groups with 58.62% overall average accuracy. Those communities that the Wilcoxon test did not generate profiles, were well characterized by the PART algorithm. The profiles generated showed the effectiveness of PART algorithm in the task of group profiling, as well as the actual relations between the two methods. All selected rules had a high accuracy rate, and 9 of the 15 rules obtained 100% of accuracy in classification.

This work is an ongoing study of group profiling in ESN. Many extensions of group profiling can be explored. In current work, we propose to understand emerging social structures based on group profiles. As future work we intended to study the adaptations of the two group profiling methods to an egocentric differentiation view, seeking minimize the imbalance among the classes and verify the suitability of the Wilcoxon test in this approach.

Acknowledgements

The authors would like to thank CNPq (Brazilian Agency) for its financial support.

References

Baradwaj, B. K. and Pal, S. (2012). Mining educational data to analyze students' performance. *CoRR*, abs/1201.3417.

Blondel, V., Guillaume, J., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10:8.

Chang, J., Boyd-Graber, J. L., and Blei, D. M. (2009). Connections between the lines: augmenting social networks with text. In *KDD*, pages 169–178.

Farzan, R. and Brusilovsky, P. (2006). P.: Social navigation support in a course recommendation system. In *In proceedings of 4th International Conference on Adaptive Hypermedia and Adaptive Web-based Systems*, pages 91–100.

Gomes, J. E. A., Prudêncio, R. B. C., Meira, L., Azevedo Filho, A., Nascimento, A. C. A., and Oliveira, H. (2013). Profiling for understanding educational social networking. *Software Engineering and Knowledge Engineering (SEKE 2013).*

Ha, S., Bae, S., and Park, S. (2000). Web mining for distance education. *Management of Innovation and Technology, 2000. ICMIT 2000. Proceedings of the 2000 IEEE International Conference on*, 2(1):715–719.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Mei, Q., Cai, D., Zhang, D., and Zhai, C. (2008). Topic modeling with network regularization. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 101–110.

Meira, L., Neves, A. M. M., and Ramalho, G. (2009). Lan house na escola: uma olimpíada de jogos digitais e educação. *Brazilian Symposium on Games and Digital Entertainment*, 8(8):150–157.

Mostow, J. and Beck, J. (2006). Some useful tactics to modify, map and mine data from intelligent tutors. *Nat. Lang. Eng.*, 12:195–208.

Newman, M. E. J. (2003). The structure and function of complex networks. *Society for Industrial and Applied Mathematics Review*, 45(2):167–256.

Romero, C. and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Syst. Appl.*, 33(1):135–146.

Sun, J., Faloutsos, C., Papadimitriou, S., and Yu, P. S. (2007). Graphscope: parameter-free mining of large time-evolving graphs. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 687–696.

Tang, L., Wang, X., and Liu, H. (2011). Group profiling for understanding social structures. *ACM Trans. Intell. Syst. Technol.*, 3:15:1–15:25.