Evolutionary Clustering and Analysis of User Behaviour in Online Forums

Donn Morrison and **Ian McLoughlin** and **Alice Hogan** and **Conor Hayes**
{*first.last*}@*deri.org*
Digital Enterprise Research Institute
National University Ireland
Galway, Ireland

Abstract

In this paper we cluster and analyse temporal user behaviour in online communities. We adapt a simple unsupervised clustering algorithm to an evolutionary setting where we cluster users into prototypical behavioural roles based on features derived from their ego-centric reply-graphs. We then analyse changes in the role membership of the users over time, the change in role composition of forums over time and examine the differences between forums in terms of role composition. We perform this analysis on 200 forums from a popular national bulletin board and 14 enterprise technical support forums.

1 Introduction

Automatic analysis of behavioural patterns in online communities has many applications for administrators, moderators and community owners. Behavioural changes of individuals or groups of influential users may help to indicate levels of community health (Angeletou and Rowe 2011), predict user churn (Karnstedt et al. 2010) and motivate changes to overall structure such as the creation, division or combination of new or existing forums.

Efforts to group users into roles according to their observed behaviour (Welser et al. 2007; Chan, Hayes, and Daly 2010; Angeletou and Rowe 2011) typically use unsupervised clustering algorithms and rely on measures such as cluster quality due to the lack of a reliable ground truth. Although these studies have provided a basis for understanding behaviour in online forums and managing the practical issues associated with such analyses, there remain many aspects of the dynamics that are not yet fully understood.

First, a study on the effect of the size of the sliding window on the results of behavioural clustering has not been carried out. Do the roles discovered using small window sizes correlate with roles discovered using larger window sizes? Second, do users of online forums maintain their behaviour over time? If so, which role(s) have longer periods of contiguous activity? Third, it is still not clear how forums vary with time in terms of their role composition. Angeletou and Rowe (2011) made progress with this in their study, but their approach used a large sliding window which makes analysis of higher frequency shifts in role composition impossible. Finally, do forums differ significantly in role composition? This question was partially addressed by Chan, Hayes, and Daly, but the approach taken may have skewed the results, as we will discuss in Section 2.

In order to address these questions, we study two sets of online forums: the popular Irish bulletin board Boards.ie and the SAP Community Network (SCN), an enterprise discussion system. We examine the extent to which individuals and groups of users maintain their behaviour over time and whether or not different forum types exhibit fundamentally different user behaviour. We also study the effect of the size of the sliding window on the resulting role composition.

2 Related work

The analysis of user behaviour in online community settings has been partially addressed in several previous works that have attempted to automate the analysis using machine learning and data mining techniques (Agrawal et al. 2003; Chan, Hayes, and Daly 2010; Kan et al. 2011; Angeletou and Rowe 2011). In such analyses, behavioural roles are typically inferred from the volume of communication and various features derived from the ego-centric network of the poster (Chan, Hayes, and Daly 2010). Agrawal et al. (2003) classified users into opposing camps of argument in Usenet newsgroups and reported better accuracy using a reply-graph partitioning algorithm than text-based classification, highlighting the effectiveness of learning from user behaviour in domains where examining the content of users exchanges may be subject to privacy restrictions, e.g. mobile phone networks.

Chan, Hayes, and Daly (2010) studied a set of 20 forums from the Boards.ie dataset and demonstrated an analysis technique that derived eight behavioural roles they believed to underlie the user base. Our work in this paper extends this study in three directions. First, Chan, Hayes, and Daly calculated and clustered ego-centric reply-graph features across all forums. This may be problematic because users may behave differently in different forums. To address this, we calculate and cluster reply-graph features for individual forums. Second, while Chan, Hayes, and Daly used a single static six month analysis window, we study how the behavioural composition of forums evolves over time using much smaller sliding windows. Finally, Chan, Hayes, and Daly relied on manually labelled clusters, while we introduce a rule-based approach that effectively maps clusters to behavioural roles.

Similarly, Angeletou and Rowe (2011) clustered users of three forums from the Boards.ie dataset into behavioural roles as a precursor to predicting community health. Using a 13 week sliding window, they used the same reply-graph features and role definitions as Chan, Hayes, and Daly. However, the authors used equal frequency unsupervised binning, a method that maps numerical variables to categorical labels, to perform the clustering and role labelling, which resulted in a high number of unclassified users that could not be analysed. Our method of evolutionary clustering assigns a prototypical role to all users.

Kan et al. (2011) examined normative user behaviour via a temporal post-reply ratio in Boards.ie and concluded that users tend to have consistent conversational behaviour over time. The difference between our work is that we consider a wider array of features and represent users as members of prototypical behavioural roles.

3 Methodology

3.1 Datasets and features

We analyse two datasets in this study. The first is the popular Irish bulletin board Boards.ie, comprising over 700 individual forums covering a wide variety of topics, 200 of which had sufficient activity (at least 10 active users per time window) for analysis. The second dataset is the SAP Community Network (SCN), an enterprise discussion system focussed on technical support of products developed by SAP, with 14 of the 33 forums having sufficient activity for analysis (as above). The analysis period for Boards.ie was the full year 2006 and for SAP was the full year 2008. The list of analysed forums is omitted due to space limitations, but selected forums are presented in Section 4 to illustrate the effectiveness of our approach.

The features used in this study are based on the egocentric reply-graphs of the users in each forum (Chan, Hayes, and Daly 2010; Rowe and Angeletou 2011). In (Chan, Hayes, and Daly 2010) fifty features were analysed and of those nine were retained following a feature correlation analysis. In this study, we make use of these nine features: *in-degree* (ind), *exponentiated in-degree* (inex), *exponentiated out-degree* (outex), *percentage of posts receiving replies* (ppr), *number of bi-directional neighbours* (bin), *number of threads with bi-directional neighbours* (thbi), *threads initiated* (th), *mean posts per thread* (mpth), *standard deviation of posts per thread* (spth). We also consider an additional feature, *forum entropy* (ent) (Rowe and Angeletou 2011). The reader is referred to (Chan, Hayes, and Daly 2010; Rowe and Angeletou 2011) for full details of these features. All features are calculated based on the reply-graph constructed for a given time window, i.e. events that occurred outside of the window are not considered. We report on the effect of the window size in Section 4.

3.2 Evolutionary clustering with K-means

K-means is a popular unsupervised clustering algorithm that converges on a local minimum by a two-step iterative process. The only parameter it requires is the number of initial centroids k. At each time step t, K-means finds a partition $\{\mathcal{V}_{1,t}, ..., \mathcal{V}_{k,t}\}$ that minimises:

$$KM = \sum_{l=1}^{k} \sum_{i \in \mathcal{V}_{l,t}} ||\vec{v}_{i,t} - \vec{\mu}_{l,t}||^2, \qquad (1)$$

where $\vec{v}_{i,t} \in \mathbb{R}^m$ is the feature vector and $\vec{\mu}_{l,t}$ denotes the centroid for cluster l. We extend K-means to evolutionary clustering by mapping centroid $\vec{\mu}_{l,t} \rightarrow \vec{\mu}_{l,t+1}$ such that K-means is initialised from where the update method for the previous time step converged (Chi et al. 2009). We demonstrate in Section 4 that this approach is a simple and effective way to perform evolutionary clustering.

3.3 User roles

The behavioural roles we defined for this study are based on those introduced by Chan, Hayes, and Daly (2010). As most forums showed a preference for four clusters,[1] we set the number of roles to be four and defined a set of decision rules to create a mapping between the clusters and roles. For consistency, we retain the primary descriptors from Chan, Hayes, and Daly and show the correspondence to the original roles in Table 1.

The decision rule listed as Algorithm 1 maps the clusters to roles using the feature means of the member elements. Prior to the mapping, the feature means are scaled into the range $[0, 1]$.

Algorithm 1 Role labelling decision rule set.

```
Input: Set of $K$ clusters $\mathcal{C}$ with cendroid means $\{$ppr, inex, outex$\}$
Output: Set of $K$ labels $\mathcal{L}$
  supporter $\leftarrow c \in \mathcal{C}$ : argmax$_c$ ppr$_c$+ inex$_c$+ outex$_c$
  $\mathcal{C} \leftarrow \mathcal{C} \backslash c$
  ignored $\leftarrow c \in \mathcal{C}$ : argmax$_c$ (1-inex$_c$) + outex$_c$
  $\mathcal{C} \leftarrow \mathcal{C} \backslash c$
  grunt $\leftarrow c \in \mathcal{C}$ : argmax$_c$ (1-outex$_c$) + (1-|inex$_c$-outex$_c$|)
  $\mathcal{C} \leftarrow \mathcal{C} \backslash c$
  elitist $\leftarrow c \in \mathcal{C}$
  $\mathcal{L} \leftarrow \{$supporter,ignored,grunt,elitist$\}$
```

3.4 Experimental setup

The analysis carried out in this study was conducted as follows. First, the reply-graph features were extracted from the forum data using the pre-selected window sizes (seven, 15 and 30 days). Next, for each forum, the features were clustered using the evolutionary extension to K-means introduced in Section 3.2 using $k = 4$ cluster centroids.

Finally, at each time step, the behavioural roles defined in Table 1 are mapped to the clusters based on the values of each clusters' feature means (see Algorithm 1). Figure 1 shows the After Hours forum from the Boards.ie dataset decomposed into the roles over the three window sizes. The role heatmap (left) shows user membership in each role (denoted by colour) at each time step. The smoothing effect as the window size increases is apparent.

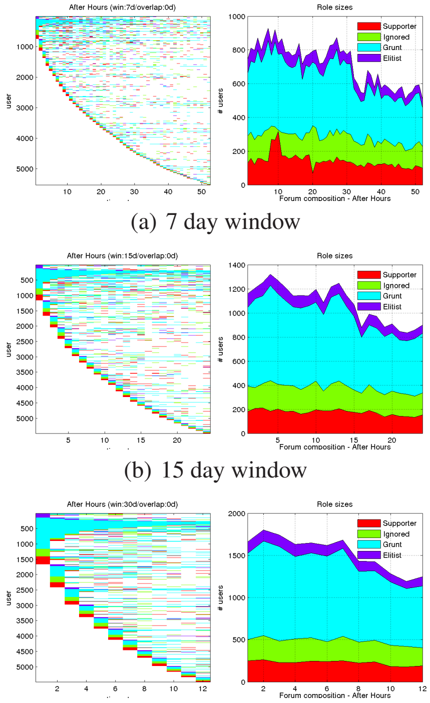
4 Results and discussion

First, we examine the effect of window size on the role composition of each forum. For each window size (seven, 15, 30

[1] Empirically determined via model selection over all forums.

Table 1: Role definitions, characteristics and correspondence to the roles defined by Chan, Hayes, and Daly (2010).

Role label	Features	Chan's roles	Characteristics
Supporter	High % posts receiving replies, high exponentiated in-degree	Supporter/popular participant/popular initiator	Stable backbone of the forum; contributes useful content that yields replies
Ignored	Low exponentiated in-degree, high exponentiated out-degree	Ignored/taciturn	Generally ignored by other users; in unmoderated forums spammers would fall into this role
Grunt	Medium-high stdev posts-per-thread, low exponentiated out-degree	Grunt/joining conversationalist	Communication with few users
Elitist	Medium-high mean posts-per-thread, medium-high exponentiated out-degree	Elitist	High communication with few users



(a) 7 day window

(b) 15 day window

(c) 30 day window

Figure 1: Evolutionary clustering of the After Hours forum from the Boards.ie dataset for three window sizes. Role heatmap (left) and forum composition (right).

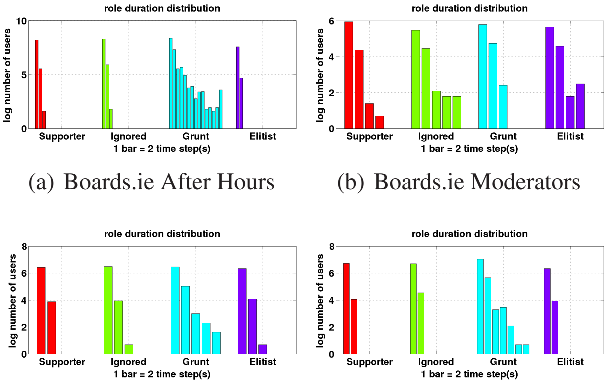
Table 2: Pearson correlation between window sizes by role (* $p < 0.05$).

(a) Boards.ie

Window sizes	Supporter	Ignored	Grunt	Elitist
7 and 15 day	0.24*	0.25*	0.34*	0.05
7 and 30 day	0.18	0.18	-0.02	-0.08
15 and 30 day	0.32*	0.28*	0.25*	0.00

(b) SAP

Window sizes	Supporter	Ignored	Grunt	Elitist
7 and 15 day	0.86*	-0.68	0.55	-0.20
7 and 30 day	0.80*	-0.04	0.62	0.11
15 and 30 day	0.79*	0.31	0.24	0.80*



(a) Boards.ie After Hours

(b) Boards.ie Moderators

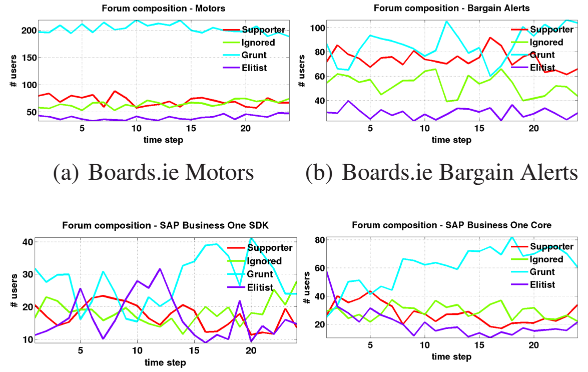
(c) SAP BO SDK

(d) SAP BO Core

Figure 2: Role duration distributions for various forums.

day), we calculated the average role composition of each forum and measured the Pearson correlation between each pair of vectors and recorded the significance values. The results are presented in Table 2 and show that the window size does have an effect on the role composition. For the Boards.ie dataset, the correlation between a seven and 30 day window is lowest, with no significance found. Correlation between seven and 15 day windows is higher, with **Supporter**, **Ignored**, and **Grunt** roles found to be significant (the **Elitist** role has no significant correlation). The same is true for the correlation between the 15 and 30 day window, which is the most correlated. Again, the **Elitist** role has almost no correlation between window sizes, suggesting that the role is highly unstable as the window size changes. For the SAP dataset, the **Supporters** role was found to be highly correlated (> 0.79, $p < 0.05$) across all window sizes. The **Ignored** role was correlated (but not significant) between seven and 15 day windows and seven and 30 day windows, but less correlated between 15 and 30 day windows. The **Elitist** role was highly correlated (0.80, $p < 0.05$) between 15 and 30 day windows. Roles derived from independent clusterings on the same window size for both datasets, on the other hand, are all highly correlated (> 0.8, $p < 0.05$).

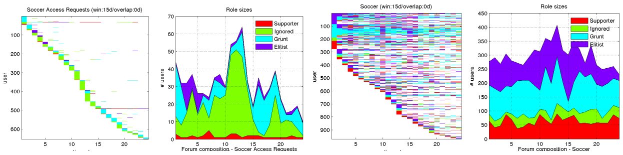
We next examine whether and to what extent users maintain their roles over time. Figure 2 shows the log distributions of role durations for a subset of forums from the Boards.id and SAP datasets. Each distribution corresponds to a role and shows the log number of users that maintained that role for n contiguous time steps, where n increases from left to right by one time step (with each bin) to the last time step (23 for the 15 day window size). For example, Figure 2 (a) shows that in the After Hours forum **Grunts** are both the most common type of user and have the longest contiguous role durations. The right-most bin stands out for that role, showing that a considerable fraction (35 users) maintain **Grunt** behaviour, uninterrupted, for the entire year. These role duration distributions, combined with the role heatmaps (such as in Figure 1), show that the majority of users maintain their roles for more than one time step, and many maintain their roles for much longer.



(a) Boards.ie Motors (b) Boards.ie Bargain Alerts

(c) SAP BO SDK (d) SAP BO Core

Figure 3: Role composition change over time for various forums.



(a) Soccer Access Requests (b) Soccer

Figure 4: Cluster heatmaps and role compositions for two related forums with very different role compositions.

The question of whether the role composition of a forum changes over time can be illustrated by Figure 3. Each plot shows the number of users in each role at each time step (15 day window). Forums such as Motors show relatively stable role compositions over time. There are no influxes of certain roles and one can expect that these would remain stable beyond the analysis window. Forums with more actively changing role compositions such as SAP BO SDK show that users change between roles, typically **Grunt** and **Elitist** for certain durations. This is confirmed by inspecting the corresponding role heatmap (not shown for these forums). Bargain Alerts also shows composition changes over time with more **Grunts** towards the end of the analysis period (which, incidentally, corresponds to the beginning of the Christmas season).

Finally, we examine the question of whether different forums differ and to what extent in their role composition. This question has already been partially answered above in the analysis of role compositions over time. We further illustrate this with some examples. Figure 4 shows two forums with a topic in common: football. The first forum, Soccer Access Requests, exists so that users can request access to the second forum, Soccer, because it is private. Users in the first forum are almost exclusively active for the window in which they issue the request, and once granted, have no need to participate further. The Soccer Access Requests has sparse activity and is composed of mainly **Ignored** users, while the Soccer forum has a wide variety of role types and a high level of activity.

5 Conclusions

This study used evolutionary clustering to group users of online forums into behavioural roles which were then analysed over time. We found that certain roles were more stable between window sizes than others (**Supporter**, **Ignored**, **Grunt** users for the Boards.ie dataset and **Supporters** for the SAP dataset). Forums with higher activity (number of users) generally had better stability between window sizes.

In some forums, considerable fractions of users maintain their roles over long periods. It was also observed that in some high activity forums, small groups of users, usually **Grunts**, maintained their behaviour for the entire analysis period. In other forums, such as Helpdesk or Soccer Access Requests, users rarely maintain their role for more than one time step due to the nature of the activity in the forum. **Grunt** users, defined as those users who communicate with few other users, were seen to be the most dominant role throughout most of the forums, especially in those that were more active, a finding supported by (Chan, Hayes, and Daly 2010).

We found that only a handful of the analysed forums, particularly the high activity forums (e.g. Motors), showed stable role compositions over time. Other forums (e.g. SAP BO Core) had compositions that changed slowly and steadily over time. Further analysis is required with knowledge of external events to determine why this may be the case. We found that generally the most active forums have similar compositions, however, some forums have very different compositions to the others (e.g. Soccer versus Soccer Access Requests) and can be explained by their nature which in turn governs the user behaviour.

Acknowledgements

This work was funded by the European Union (EU) and Science Foundation Ireland (SFI) in the course of the projects ROBUST (EU grant no. 257859) and CLIQUE Strategic Research Cluster (SFI grant no. 08/SRC/I1407), and LION-2 (SFI grant no. SFI/08/CE/I1380).

References

Agrawal, R.; Rajagopalan, S.; Srikant, R.; and Xu, Y. 2003. Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th international conference on World Wide Web*, 529–535. ACM.

Angeletou, S., and Rowe, M. 2011. Modelling and analysis of user behaviour in online communities.

Chan, J.; Hayes, C.; and Daly, E. M. 2010. Decomposing Discussion Forums and Boards Using User Roles. In *International AAAI Conference on Weblogs and Social Media*, 215–218.

Chi, Y.; Song, X.; Zhou, D.; Hino, K.; and Tseng, B. L. 2009. On evolutionary spectral clustering. *ACM Transactions on Knowledge Discovery from Data* 3(4):1–30.

Kan, A.; Chan, J.; Hayes, C.; Hogan, B.; Bailey, J.; and Leckie, C. 2011. A Time Decoupling Approach for Studying Forum Dynamics. *World Wide Web Internet And Web Information Systems* In press:1–24.

Karnstedt, M.; Rowe, M.; Chan, J.; Alani, H.; and Hayes, C. 2010. The Effect of User Features on Churn in Social Networks. *Human Factors*.

Rowe, M., and Angeletou, S. 2011. Predicting discussions on the social semantic web. *Web: Research and Applications*.

Welser, H.; Gleave, E.; Fisher, D.; and Smith, M. 2007. Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure* 8(2):564–586.