Decomposing Discussion Forums and Boards Using User Roles

Jeffrey Chan and **Conor Hayes**
Digital Enterprise Research Institute
NUI Galway
Ireland

Elizabeth M. Daly
IBM, Dublin Software Lab
Ireland

Abstract

Discussion forums are a central part of Web 2.0 and Enterprise 2.0 infrastructures. The health and sustainability of forums is dependent on the information exchange behaviour of its contributors, which is expressed through online conversation. The increasing popularity and importance of forums requires a better understanding and characterisation of communication behaviour so that forums can be better managed, new services delivered and opportunities and risks detected. In this paper, we present an empirical analysis of user communication roles in a medium-sized bulletin board and we analyse the composition of several forums in terms of these roles, demonstrating similarities between forums based on underlying user behaviour rather than topic.

1. Introduction

Discussion forums are a central part of Web 2.0 and Enterprise 2.0 infrastructures. The increasing popularity and importance of forums in supporting online communities requires a better understanding and characterisation of communication behaviour so that forums can be better managed, new services delivered and opportunities and risks detected. The health and sustainability of a forum is dependent on the information exchange behaviour of its members, which is expressed through conversations.

While forums haves been the topic of several studies (Feng et al. 2006), their constitution in terms of user roles has up until now remained unexplored. Roles might include a topic instigator, who tends to initialise popular threads, or a taciturn contributor who tends to ask questions but only engages in limited conversion. Such analysis will enable host organizations to assess the health of their forums, and make decisions on resources such as moderatorship or additional support.

Manual role identification of large-scale data is time consuming and infeasible. This paper contributes an automated forum profiling technique to capture and analyse user behaviour which is empirically evaluated using a medium sized national discussion board dataset. Our analysis found that forums are typically composed of eight behaviour types such as 'popular initiators', 'grunts' and 'taciturns'.

In Section 2, we present related work. Section 3 describes the data set and its representation. In Section 4, we describe how we identify user roles and the analysis of forums based on these roles. Section 5 presents a comparative analysis of the role composition of forums. Section 6 concludes this paper and discusses future work.

2. Related Work

The feature-based approach of (Fisher, Smith, and Welsher 2006) analysed a user's ego-centric network and the out-degree distribution along with visual representations of the network. However, they did not try to fit a distribution to the out-degree plots which would enable the automated incorporation of out-degree distribution features. (Welser et al. 2008) extended the ego-centric network analysis to determine roles such as technical editor and substantive expert in Wikipedia. These approaches had a significant element of manual anlysis and thus are not easily scaled. (Ellison, Steinfield, and Lampe 2007) used regresion to determined the dependency between relationship strength on Facebook and social capital, determined using a survey. (Barash et al. 2009) used role- and context-related features to classify whether messages are factual (e.g., technical) or relational (e.g., opinion and support). Unlike our work, both these papers are concerned with supervised learning of the relation between a target and a set of features. However, we are interested in determining discrete groupings of features to discover common roles. (Himelboim, Gleave, and Smith 2009) identified social leaders in political forums using ratios based on the amount of replies to posts and threads initiated by a user. However, we found that these ratios were most likely a result of the scale-free characteristics of the interaction graph.

In regular equivalence (Lerner 2004; Wasserman and Faust 1994), two users play the same role if they are connected to the same types of users. However, it is difficult to incorporate non-binary features (like number of replies between two users) into the role equivalence model and techniques. Finally, a number of previous works have focused on analysing question and answer style discussion forums (Feng et al. 2006; Hong and Davison 2009). Similar to our work, (Adamic et al. 2008) analysed the communication graph of the forums of a Q&A website and used features such as thread length, amount of replies to questions,

in and out degree, etc. to classify a forum. Unlike our work, they do not break down a forum into the composition of user roles.

In summary, all the presented related work either used manual methods to analyse user roles and forums, which are not scalable, limited to unweighted relations (role equivalence), or are focused on Q&A forums only. In this paper, we present our forum analysis approach that is automated and scalable to larger forums, can analyse weighted features, and can be applied to any type of discussion forums.

3. Data Sets

Boards.ie is the largest general topic discussion board in Ireland. In the last 12 months, there were 596 forums, 244850 threads, 75400 users and over 4.3 million posts. To represent the communication interaction between users, we model the interaction as a weighted, directed graph. Each vertex represent a user in a forum, and a directed edge exists from user v_i to user v_j if user v_i has replied to a post of user v_j in thread t_k in the forum. We also associate the number of posts between two users as the edge weight. Note that from this definition, multi-edges can exist between two users, with each directed edge representing reply-behaviour from one user to another in a particular thread. We call this graph the *reply graph*. The *collapsed reply graph* aggregates all the multi-edges into a single edge, with the weight of the resultant edge being the sum of the weights of the multi-edges. The reply graph is used to analyse reciprocity of communications between users and which types of users are communicating. To demonstrate the profiling technique, in this paper we focus our analysis to 20 different forums from boards.ie from the period 01/07/2006 to 31/12/2006, inclusive. The forums represent a range of topics from discussion to technical to advertisement. The method is general and can be applied to any number of forums.

4. Forum Composition Approach

In this section, we explain the features we used. Some of the features are discriminating within a forum, while others are useful for analysing users across forums. We analysed approximately 50 different features, but many of these were highly correlated with each other, hence redundant for grouping purposes.

Structural Features Structure features provide an indication of the communication between users and can be derived from the properties of the unweighted, directed graph. A user can be characterised by the interactions of his neighbours, which we study by analysing his ego-centric network (Wasserman and Faust 1994). We found the ego-centric networks follow a power-law distribution - further confirmed by the low clustering coefficient of all ego-centric networks. The in- and out- degree distributions of neighbours in the ego-centric networks followed a power law distribution, which we represented by its exponent (**in-degree exponent**, **out-degree exponent**).

Reciprocity Features The feature **% of bi-directional neighbours** represents the percentage of the neighbours of a user where there is both in and out edges (i.e. they have replied to each other). In addition, we analysed the percentage of *threads* in which a user has reciprocal communication with at least one other user (the two users have replied to each other's post in the thread). A user can have a low percentage of bi-directional neighbours but a high percentage of threads in which there is at least one reciprocal communication.

Persistence Features We measure the mean and standard deviation of the posts per thread (**average post/thread**, **std. dev. post/thread**).

Popularity Features The more popular a user, the more likely are they to be replied to. **in-degree %** is the ratio of a user's in-neighbours compared to all users that have replied to someone else. **% Posts Replied** measures the percentage of posts where there is at least one reply to the user. A user can have many repliers but only a low percentage of her posts actually receive replies.

Initialisation Features **initiated %** measures what percentage of threads are initiated by a user. It can distinguish users who initiate many threads from those that just replies.

We have presented nine different features used for grouping users into common roles. Next we describe how we perform the grouping to find the common roles.

4.1 User Role Discovery Approach

Using principle component analysis to analyse the features, we found that the the amplitude of the largest principal component constituted more than 95% of the variance in the features, and the size of the ego-centric networks was the dominant feature in the largest component. Hence, we use the size of the ego-centric networks as our feature to partition the users into the three bands. We discard the lowest band, which consists one-post users, and the middle band, which does not have enough neighbours to have an accurate power law exponent fit. Using agglomerative hierarchical clustering, we cluster the feature profile data of the remaining top band users from all forums. To determine the optimal number of clusters, we used five different validation techniques: Rand, Silhouette, RS, Root mean square and DB Index (Handl, Knowles, and Kell 2005). We found that the optimal number of clusters was either 8, 13, 15 or 21. After manual inspection, we selected 8 and 15 as the best numbers of clusters. Each cluster approximately corresponds to one user role type. The average value of the nine features and the number of users in each cluster are used to build a quantitative description of the clusters/user role types. We manually grouped the 15 types into 8 role categories. While this process was informed by the *k*=8 partitioning, we also noted that this partitioning would not have discovered all of the roles summarised in Table 1. In the next section, we describe how we classify forums based on their role composition.

5. Forum Composition in Boards.ie

In this section, we show that we obtain different and unexpected groupings of the 20 selected forums based on role

Name	Clusters	Comments
Joining Conversationalists	1, 2	No initialisation. High levels of communications with a relatively small set of users.
Popular Initiators	3, 13	Very high levels of thread initialisation, coupled with relatively high popularity (high in-deg %).
Taciturns	5, 6	Very low reciprocity, volume of communication and few neighours suggest limited conversation with a few users. The main difference between clusters 5 and 6 is their exponents, suggesting they communicate with different types of neighbours.
Supporters	4, 7	Relatively middle of the road statistics, suggesting the users form the backbone of the forums. Difference between clusters 4 and 7 is the amount of communications.
Elitists	9	Characterised by very low percentage of neighbours with bi-directional communications but high percentage for bi-directional threads. Combined with low in-deg percentage, these users prefer to carry on conversation with a very small set of users.
Popular Participants	8, 12, 14	Do not initiate much threads, unlike the popular initiators, but are involved with a large percentage of users on forums. They can be considered a cross between joining conversationalist and popular initiators. The difference between clusters 8 and 12 is the volume of communications.
Grunts	10, 11	Low volumes of communications to a few users. Different from taciturns by the relatively higher levels of reciprocity.
Ignored	15	Very low percentage of posts get replied to

Table 1: Summary of the common user roles.

compositions. Using an unweighted Euclidean distance, we cluster the forums into groups (see Table 2)

For reasons of space, Figure 1 only shows the role composition for 8 of the 20 forums analysed. Visually, we can see some forums are distinctly different from the others, such as the personal issues forum. But there are also some forums that have similar compositions, such as the accommodation and politics forums.

Id	Forums
1	Personal Issues
2	Christianity
3	Paranormal
4	Thunderdome
5	Overclocking
6	Weather
7	Windows, Development, Humanities, Accommodation, Politics
8	Travel, Gigs & Events
9	Soccer, Poker, UCD, Playing Instruments
10	Martial Arts, TCD, Tournaments & Events

Table 2: Forum groupings.

The singleton clusters numbered 1 to 6 have very different composition to all other forums. For example, the *taciturn* role makes up 95% of all users in the Personal Issues forum (grouping 1). This suggests that, despite its name, there is little dialogue happening. Grouping 2 (the Christianity forum) has a strong component of popular initiators, suggesting that a few users regularly initiate threads that subsequently generate discussion (large percentage of popular participants and supporters). Grouping 6, the weather forum, is also strongly constituted by popular initiators and popular participants, but it has a larger portion of grunts and supporters than the Christianity forum, suggesting that lengthy discussion is not as widespread as in the Christianity forum.

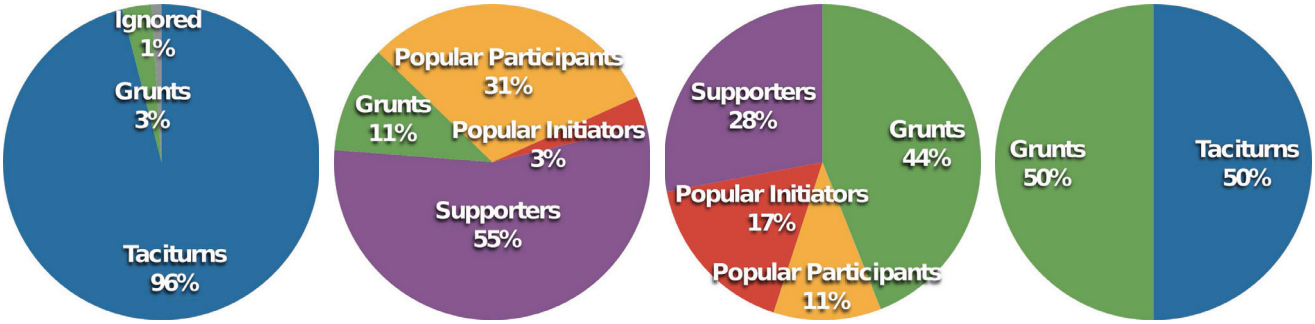
Groupings 7, 9 and 10 consists of four forums each. However, the crucial difference between the three groupings is the relative proportions of grunts, popular participants, supporters and taciturns. For reasons of space, we focus only on grouping 7. In comparsion to the Weather or Christianity forums, the forums in grouping 7 appear to be much less social, dominated by taciturn and grunt roles. At face value this may suggest that these forums are not functioning well. However, this may not be the case as two of these forums are technical support forums (Windows and Development) where a question-and-answer format may be the most usual form of communication. Similarly, the accomodation forum tends to be made up of personal notices seeking or advertising accomodation rather than discussion. The humanities forum has a clique of highly social elitists and a sizeable number of supporters. As with the politics forum, it is difficult to say without further analysis whether these two forums are functioning to the satisfaction of their participants.

6. Conclusion

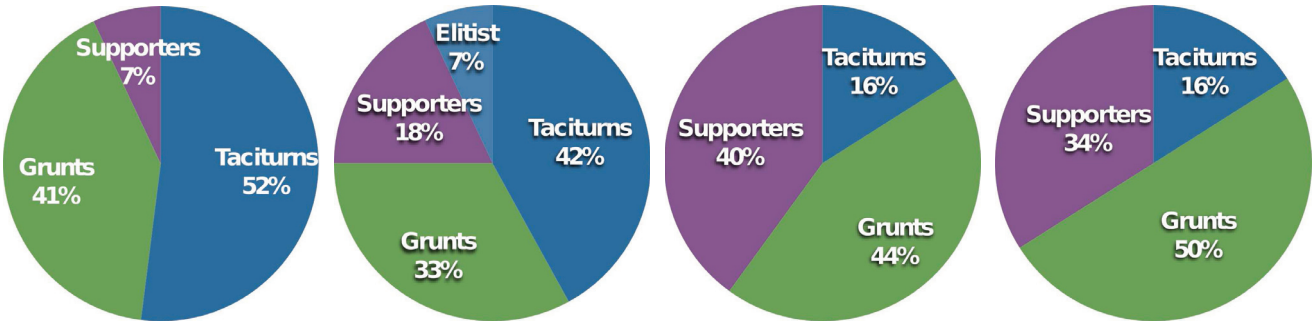
In this paper, we have presented a novel method of analysing forums in terms of the roles played by users. We used nine different features to profile the user roles. We then applied a two stage clustering approach to group the users of the forums into 15 groups and eight roles. Using these roles, we describe the forums based on their role composition. We showed how the forums can be clearly compared, analysed and grouped based on composition. In further work, we plan to analyse the role composition across time. And to understand what are the composition norms for differnet types of forums. In addition, we would like to extend our role composition technique to other online domains such as weblogs.

7. Acknowledgements

This work is supported by the Science Foundation Ireland (SFI) under CLIQUE Strategic Cluster, grant number 08/SRC/I1407. We would like to thank John Breslin for



(a) 1. Personal Issues. (b) 2. Christianity (c) 6. Weather. (d) 7a. Windows.



(e) 7b. Development. (f) 7c. Humanities. (g) 7d. Accommodation. (h) 7e. Politics.

Figure 1: The user roles composition of the 20 forums. The colour scheme for the clusters is consistent across the forums, so for example, cluster 1 has the same colour across the pie charts.

kindly providing the Boards.ie data and Ron Kass of Boardtracker.com for insightful discussions and board statistics.

References

Adamic, L. A.; Zhang, J.; Bakshy, E.; and Ackerman, M. S. 2008. Knowledge sharing and yahoo answers: everyone knows something. In *Proceeding of the 17th international conference on World Wide Web*, 665–674. New York, NY, USA: ACM.

Barash, V. D.; Smith, M.; Getoor, L.; and Welser, H. T. 2009. Distinguishing knowledge vs social capital in social media with roles and context. In *Proceedings of the ICWSM 09*.

Ellison, N. B.; Steinfield, C.; and Lampe, C. 2007. The benefits of facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication* 12(4).

Feng, D.; Shaw, E.; Kim, J.; and Hovy, E. 2006. Learning to detect conversation focus of threaded discussions. In *Proceedings of HLT/NAACL 2006*, 208–215. Morristown, NJ, USA: Association for Computational Linguistics.

Fisher, D.; Smith, M.; and Welsher, H. T. 2006. You are who you talk to: detecting roles in usenet newsgroups. In *Proceedings of the 39th Annual Hawaii International Conference*, volume 3, 59–68.

Handl, J.; Knowles, J.; and Kell, D. 2005. Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21(15):3201.

Himelboim, I.; Gleave, E.; and Smith, M. 2009. Discussion catalysts in online political discussions: Content importers and conversation starters. *Journal of Computer-Mediated Communication* 14(4):771–789.

Hong, L., and Davison, B. D. 2009. A classification-based approach to question answering in discussion boards. In *Proceedings of ACM SIGIR conference on Research and development in information retrieval*, 171–178. New York, NY, USA: ACM.

Lerner, J. 2004. Role assignments. In *Network Analysis*, 216–252.

Wasserman, S., and Faust, K. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1 edition.

Welser, H.; Kossinets, G.; Smith, M.; and Cosley, D. 2008. Finding social roles in wikipedia. In *Presented at the Annual Meeting of the American Sociological Association Annual Meeting*, 1–11.