What Catches Your Attention? An Empirical Study of Attention Patterns in Community Forums

Claudia Wagner
Institute of Information and Communication Technologies
JOANNEUM RESEARCH
Graz, Austria
claudia.wagner@joanneum.at

Matthew Rowe
Knowledge Media Institute
The Open University
Milton Keynes, United Kingdom
m.c.rowe@open.ac.uk

Markus Strohmaier
Knowledge Management Institute
Graz University of Technology
Graz, Austria
markus.strohmaier@tugraz.at

Harith Alani
Knowledge Media Institute
The Open University
Milton Keynes, United Kingdom
h.alani@open.ac.uk

Abstract

Online community managers work towards building and managing communities around a given brand or topic. A risk imposed on such managers is that their community may die out and its utility diminish to users. Understanding what drives attention to content and the dynamics of discussions in a given community informs the community manager and/or host with the factors that are associated with attention. In this paper we gain insights into the idiosyncrasies that individual community forums exhibit in their attention patterns and how the factors that impact activity differ. We glean such insights by using logistic regression models for identifying seed posts and explore the effectiveness of a range of features. Our findings show that the discussion behaviour of different communities is clearly impacted by different factors.

Introduction

Social media applications such as blogs, video sharing sites or message boards allow users to share various types of content with a community of users. The different nature and intentions of online communities means that what drives attention to content in one community may differ from another. For example, what catches the attention of users in a question-answering or a support-oriented community may not have the same effect in conversation-driven or event-driven communities. In this paper we use the number of replies that a given post on a community message board yields as a measure of its attention and explore factors that impact the attention level a post gets in certain community forums.

Through an empirical study of attention patterns in 10 different forums on the Irish community message board Boards.ie[1], we analysed how attention is generated in different community forums. Our study was facilitated through a classification experiment which aims to identify seed posts - i.e. thread starter posts on a community message board that got at least one reply - and the use of five distinct feature sets - *user*, *focus*, *content*, *community* and *post title* features. We find interesting differences between these communities in terms of what drives users to reply to thread starters initially. Our work is relevant for researchers interested in behavioural analysis of communities and analysts and community managers who aim to understand the factors that are associated with attention within a community.

[1]http://www.boards.ie

Dataset: Boards.ie

In this work, we analysed data from an Irish community message board, Boards.ie, which consists of 725 community forums ranging from communities around specific computer games or spiritual groups to communities around general topics such as films or music. For our analysis we used all data published in the year 2006. Table 1 describes the properties of the dataset.

Since our goal was to uncover the idiosyncrasies of individual community forums and the deltas between them, we selected 10 distinct forums for analysis of their attention patterns. These forums were selected by computing 5 statistics using data from 2005 (*average post count, average number of users, average number of replies, average number of seeds* and *average number of non-seeds per forum*), plotting each community in a PCA space and then selecting forums that appeared away from one another in the space. Table 2 provides a brief description of the selected community forums.

Feature Engineering

Understanding what factors drive reply behaviour in online communities involves defining a collection of features and then assessing which are important for identifying seed posts. We defined the following five feature groups: *User features* describe the author of a post via his/her past behaviour, while *focus features* measure the topical concentration of posts by an author. *Post features* capture characteristics of a post, while *title features* focus on the title of a post itself and identify attributes that the title should contain in order to start a discussion. *Community features* describe relations between a post or its author and the community with which the post is shared. Table 3 provides a brief description of the features we used and relates each feature with a

Table 1: Description of the Boards.ie dataset

Posts	Seeds	Non-Seeds	Replies	Users
1,942,030	90,765	21,800	1,829,465	29,908

Table 2: Overview of selected community forums

ID	Name	Description
7	After hours	General discussion forum with the highest level of activity on the platform.
9	Computers and Technology	Computer support-oriented forum containing posts enquiring about issue resolution.
552	Wanted General	Forum where users state items and products that they would like which other users could provide.
483	Cuckoo's Nest	Conversation forum for liberally minded individuals.
47	Motors	Contains posts related to motoring spanning topics such as new cars, purchasing advice and general motoring discussion.
11	Flight Simulator General	Community for discussions about the video game Flight Simulator.
556	Wanted Tickets	Forum for users to state their needs for event tickets, ranging from sports through to music concerts.
468	TCD	Forum for discussions related to Trinity College Dublin (TCD), one of the largest universities in Ireland.
411	Mobile Phones and PDAs	Contains discussions related to mobile phone issues and portable devices that are emerging on the market. Often contains support requests and allows users to resolve problems they are having.
453	Flight Simulator Discs	Forum for the exchange and sale of computer discs for the video game Flight Simulator.

feature group.

For each thread starter post we computed the features by taking a 6-month window prior to when the post was made. That means, we used all the author's past posts within that window to construct the necessary features - i.e. constructing a social network for the user features, assessing the forums in which the posts were made for the focus features and inferring topic distributions per user and month based on the content of posts he/she authored within the previous 6 month. For the features that relied on topic models, we first trained a Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) model which we use later for inferring users' topic distributions. For training the LDA model we aggregated all posts authored by one user in 2005 into an artificial user document and chose the default hyperparameters ($\alpha = 50/T$, $\beta = 0.01$ and $T = 50$) which we optimised during training. We used this model to infer the monthly average topic distributions (averaged over 10 independent runs of a Markov chain) of users who authored at least one post in 2006 based on all posts they authored within the last 6 months. We use monthly-increments for scalability.

Experimental Setup

Our experiment sought to identify the factors that were associated with discussions in different communities. To that end, we conducted binary classification experiment using a logistic regression model and the features as described in the previous section. For each forum, we divided the forum's dataset into a training/testing split using an 80/20% split, trained the logistic regression model using the former split and applied it to the latter. We tested each of the five feature sets in isolation - i.e. user, focus, post, community and title - such that the model was trained using only those features, and then tested all the features combined together. The best performing model was then chosen and the coefficients of the logistic regression model were inspected to detect how the features were associated with seed posts, thereby identifying the factors that impact reply behaviour of users in different community forums.

To assess how well each model performed, we measured the Area Under the ROC Curve (AUC). A curve that maximises the AUC, and therefore achieves $AUC = 1$, is optimal.

Results: Seed Post Identification

Comparing the AUC values of different forums in Table 4 reveals interesting differences between communities and corroborates our hypothesis that the reply behaviour of users in different communities is impacted by different factors. While content features are most important for community forum 411 (Mobile phones and PDAs), user features are most important for the communities around forum 453 (Flight Simulator Discs) and 483 (Cuckoo's nest). That means that in forum 411 it mainly depends on a post and its characteristics whether the post gets replies or not, while in forum 453 and 483 posts are far more likely to get replies if they were authored by certain types of users. For the communities around forum 556 (Tickets wanted), 552 (Wanted) and 11 (Flight Simulator General), which all have relatively low discussion levels (i.e. many posts get no replies), community features were most important for predicting which post will get replies. It suggests that in those communities only posts and/or users which fit into the community and/or contribute to the community will get replies. Finally, for the communities 7 (After Hours), 9 (Computers and Technology), 468 (TCD) and 47 (Motors), a classifier based on all features performed best in differentiating between posts which get replies and posts which do not stimulate any discussion.

To gain deeper insights into the factors which impact users' reply behaviour, we further analysed the coefficients of the logistic regression model which indicate the features' influence on the probability of a post getting replies. In the following we only discuss statistical significant coefficients. For example, when further analysing the Mobile phones and PDAs community, for which content factors seem to play a crucial role, we noted that in this community posts which have a higher polarity ($c = 3.14$) and are therefore more positive are far more likely to get replies. This community seems to be mainly driven by content factors, while characteristics of authors or relations between authors and the rest of the community play a minor role. Community 9 (Computers and Technology) seems to have a supportive purpose. Posts are far more likely to get replies if titles contain question marks ($c = 0.528$), articles ($c = 0.0211$) and negated words ($c = 0.0581$) and if the post's content has

Table 3: Overview of the features and their group memberships.

Group	Name	Description
User	User Account Age	Measures the length of time that the user has been a member of the community.
User	Post Count	Measures the number of posts that the user has made.
User	Post Rate	Measures the number of posts made by the user per day.
User	In-degree	Measures the number of incoming connections to the user.
User	Out-degree	Measures the number of outgoing connections from the user.
Focus	Forum Entropy	Measures the forum focus of a user via the entropy of a user's forum distribution. Low forum entropy would indicate high focus.
Focus	Forum Likelihood	Measures the likelihood that the user will publish a post within a forum given the past forum distribution of the user.
Focus	Topic Entropy	Measures the topical focus of a user via the entropy of a user's topic distributions inferred via the posts he/she authored. Low topic entropy would indicate high focus.
Focus	Topic Likelihood	Measures the likelihood that the user will publish a post about certain topics given his/her past topic distribution. Therefore, we measure how well the user's language model can explain a given post by using the likelihood measures: $$likelihood(p) = \sum_{i=0}^{N_p} \ln P(w_i \mid \hat{\phi}, \hat{\theta}) \quad (1)$$ N_p refers to the total number of words in the post, $\hat{\phi}$ refers to the word-topic matrix and $\hat{\theta}$ refers to the average topic distribution of a user's past posts. The higher the likelihood for a given post, the greater the post fits to the topics the user has previously written about.
Focus	Topic Distance	Measures the distance between the topics of a post and the topics the user wrote about in the past. We use the Jensen-Shannon (JS) divergence to measure the distance between the user's past topic distribution and the post's topic distribution. The lower the JS divergence, the greater the post fits the topics the user has previously written about.
Post	Post Length	Measures the number of words in the post.
Post	Complexity	Measures the cumulative entropy of terms within the post, using the word-frequency distribution, to gauge the concentration of language and its dispersion across different terms.
Post	Readability	This feature gauges how hard the post is to parse by humans by using Gunning fog index (Gunning 1952) which uses average sentence length (ASL) and the percentage of complex words (PCW): $0.4 * (ASL + PCW)$.
Post	Referral Count	Measures the number of hyperlinks within the post.
Post	Time in day	The number of minutes through the day from midnight that the post was made. This feature is used to identify key points within the day that are associated with seed or non-seed posts.
Post	Informativeness	Measures the novelty of the post's terms with respect to other posts. We derive this measure using the Term Frequency-Inverse Document Frequency (TF-IDF) measure.
Post	Polarity	Assesses the average polarity of the post using Sentiwordnet.[2]
Community	Topical Community Fit	Measures how well a post fits the topical interests of a community by estimating how well the post fits into the forum. We measure how well the community's language model can explain the post by using the likelihood measure which is defined in equation 1, where $\hat{\theta}$ refers to the average topic distribution of posts that were previously published in that forum. The higher the likelihood of the post, the better the post fits to the topics of this community forum.
Community	Topical Community Distance	Measures the distance between the topics of a post and the topics the community discussed in the past. We use the Jensen-Shannon (JS) divergence to measure the distance between a community's past topic distribution and a post's topic distribution. The lower the JS divergence, the greater the post fits the topical interests of the community.
Community	Evolution score	Measures how many users of a given community have replied to a user in the past, differing from *in-degree* by being conditioned on the forum. Theories of evolution (McKelvey 1997) suggests a positive tendency for user A replying to user B if A previously replied to B.
Community	Inequity score	Measures how many users of a given community a user has replied to in the past, differing from *out-degree* by being conditioned on the forum. Equity Theory (Adams 1965) suggests a positive tendency for user A replying to user B if B previously replied more often to A than A to B.
Title	Length	Number of words in the title of the post.
Title	Questionmark	Measures the absence or presence of a question-mark in the title.
Title	Linguistic Dimension	Measures the proportion of words per linguistic dimension using LIWC (Linguistic Inquiry and Word Count) (Tausczik and Pennebaker 2010) which categorises 2300 words or word stems into over 70 linguistic dimensions. Rather than using all 70 dimensions we chose five evocative dimensions for our analysis and derived a feature for each one: human terms (e.g. adult, baby), anger (e.g. hate, loathe), sexual (e.g. horny, love), article (e.g. a, an) and negate (e.g. no, not).

Table 4: Area under the ROC curve (AUC) for different forums when performing seed post identification

Forum	User	Focus	Content	Commun'	Title	All
7	0.612	0.660	0.661	0.536	0.522	**0.711**
9	0.556	0.590	0.559	0.463	0.568	**0.631**
552	0.434	0.469	0.510	**0.532**	0.518	0.502
483	**0.918**	0.890	0.415	0.765	0.530	0.700
47	0.573	0.542	0.631	0.490	0.548	**0.687**
11	0.596	0.539	0.578	**0.604**	0.410	0.603
556	0.434	0.545	0.624	**0.683**	0.465	0.552
468	0.597	0.582	0.473	0.442	0.570	**0.601**
411	0.469	0.468	**0.526**	0.396	0.497	0.489
453	**0.678**	0.602	0.509	0.574	0.585	0.612

high complexity ($c = 0.988$), and therefore uses more expressive language. Outsiders, i.e. users which seem to be rather new to the topic they are writing about (high topic distance $c = 0.970$) and which are not really focused on this particular forum (high forum entropy $c = 0.163$), are more likely to get replies. Interestingly, long titles (title length $c = -0.0109$) and long posts ($c = -0.0103$) have a negative impact on posts getting replies in such support oriented forums. Users who replied to many others (higher outdegree $c = -0.0216$) in the past are also less likely to get replies. Similarly the community around forum 47 (Motors) also seems to have a supportive purpose where content is an important factor for anticipating the start of discussions. Posts which fit into the community (high topical community fit $c = 0.0758$), whose title contains question marks ($c = 0.0554$) and whose content contains a wider vocabulary of terms (high complexity $c = 0.719$) are more likely to catch the attention of this community.

Communities oriented around a very specific subject such as the community in forum 468 (Trinity College Dublin) are more likely to reply to users who are new to the platform (lower user account age $c = -1.58E^{-5}$) and the topic of community's interest (high topic distance $c = -3.53$). The more engaged a user is in a forum (high forum likelihood $c = 0.192$) and the more positive his/her post is (high polarity $c = 3.968$) the more likely he/she will catch the attention of this community. This suggests that naivety of the user plays a role, where a new or prospective student could be asking the community for information about the university. Communities which are oriented around a more general subject, such as the one around forum 7 (After Hours) also require users to engage in a forum (high forum likelihood $c = 6.94$) but do not require them to only focus on one community (high forum entropy $c = 0.379$) in order to get replies. New users (high topic distance $c = 2.00$) which have a topical focus (low topical entropy $c = -0.515$) are likely to get replies. Further, short posts ($c = -0.0117$) which have high complexity ($c = 0.797$) are as well more likely to attract the attention of this community.

Conclusions and Future Work

In this paper, we have presented work that identifies attention patterns in community forums and shows how such patterns differ between communities. Our findings demonstrated that different community forums exhibit interesting differences in terms of how attention is generated. Our results suggest understanding the purpose and nature of a community, including the specificity of its subject, seems to be crucial for identifying the right features to anticipate community behaviour. Communities that seem to have a partly supportive purpose (such as community 9 and 47) tend to be content driven and such communities are more likely to reply to users who are new to the area, not greatly involved in the community and who are seeking help by publishing a post which is about a topic which fits in the community. Communities around very specific subjects (such as the community 468) tend to reply to users who are new to the community and focussed, while communities around more general subjects such as the After Hours community (7) do not have this requirement. In communities that lack specificity everyone can participate, but posts are required to be rather short in order to minimise effort while still containing distinct terms in order to attract attention. We also note that for support-oriented communities there are common patterns in the inclusion of a question-mark and complexity of the language used - requiring an wider vocabulary of terms.

Although our work is limited to a small number of communities on one message board platform, Boards.ie, it uncovers an interesting problem: the problem of identifying the context in which attention patterns may occur. Our results show that the attention patterns of different communities are impacted by different factors and therefore suggest that these patterns may only be valid in a certain context and that the existence of global, context-free attention patterns is highly questionable. Our previous work in (Rowe, Angeletou, and Alani 2011) focussed on identifying global attention patterns and suggested that the initial reply behaviour of communities on Boards.ie tends to be driven by content-factors while our findings show that this is only true for certain types of communities. Our future work will explore this avenue by comparing similar communities for the existance of similar attention patterns.

Acknowledgment

Claudia Wagner is a recipient of a DOC-fForte fellowship of the Austrian Academy of Science. The work of Matthew Rowe and Harith Alani was supported by the EU-FP7 project Robust (grant no. 257859).

References

Adams, J. 1965. Inequity in social exchange. *Adv. Exp. Soc. Psychol.* 62:335–343.

Blei, D. M.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *JMLR* 3:993–1022.

Gunning, R. 1952. *The Technique of Clear Writing*. McGraw-Hill.

McKelvey, B. 1997. Quasi-natural organization science. *Organization Science* 8(4).

Rowe, M.; Angeletou, S.; and Alani, H. 2011. Anticipating discussion activity on community forums. In *The Third IEEE International Conference on Social Computing*.

Tausczik, Y. R., and Pennebaker, J. W. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29(1):24–54.