

On the Correlation between Topic and User Behaviour in Online Communities

Erik Aumayr and Conor Hayes

Insight Centre for Data Analytics

National University of Ireland, Galway

E-mail: {erik.aumayr, conor.hayes}@insight-centre.org

Abstract

In the advancement of a better understanding of user behaviour on the Internet, clustering approaches are used as a means to group and categorise online communities. Although clustering approaches indicate similarity between communities (e.g. through an a priori defined distance function based on user behaviour features) they do not show what causes the differences between the types of communities. In this paper, we study the effects that certain types of topics may have on user behaviour. We cluster 29 online communities of the Stack Exchange platform, and we show the correlation between the clustered user behaviour and the three latent topic properties Accessibility, Sociability and Controversy.

1 Introduction

Wherever people meet and interact with each other they form communities based on common interests. The definition of community varies from situation to situation. In a medieval example, settlements and villages were typical communities, offering combined efforts of support on the inside and protection towards the outside. In the modern world, where almost half of all people interact on the Internet¹, they form online communities where they discuss topics of interest and provide support to their members. There are many examples of online communities to date: social media sites (e.g. Facebook and Twitter), mailing lists (e.g. Google Groups), online discussion forums (e.g. Reddit), and Q&A sites (e.g. Stack Exchange). Indeed, online communities are an important means in today's world for people and businesses to share knowledge and provide support much faster and in a much larger scale.

In order to make sense of the large amount of different types of online communities, researchers often employ clustering approaches. Depending on the extracted features and the distance function of the clustering algorithm, we can then group and categorise the communities by the different ways users interact. As we will see in Section 3, that provides us with an understanding of different aspects of communities, for example, which communities grow faster. What the clustering does not tell us is why users behave differently. In this work, we study the effects that certain kinds

of topics might have on user behaviour. We define three latent topic properties: Accessibility, Sociability and Controversy; and show that they are correlated with the user behaviour in community clusters.

2 Related Work

In order to understand the differences between online communities, so-called taxonomies or typologies have been employed that group the communities in a top-down fashion according to some high-level aspects, such as *relationship orientation* (social or professional) and *establishment* (commercial, non-profit and government) (Porter 2004), or the community's *purpose* and *platform* (Lazar and Preece 1998). Although communities can be categorised with these and similar dimensions, for example into *discussion or conversation* communities, and *task- and goal-oriented* communities (Stanoevska-Slabeva 2002), the authors do not show whether these high-level dimensions are reflected in measurable user behaviour. For instance, the users in a discussion-based community may or may not produce longer threads than the ones in a goal-oriented community.

Other works use a bottom-up approach and cluster the communities by measurable user behaviour. Chan et al. use a set of 9 features that are mainly focussed on the user-graph, as part of their study of user roles in the Irish general discussion forums of Boards.ie (Chan, Hayes, and Daly 2010). It is notable that they remove low-activity users as they introduce noise in their user role analysis, although these users make up the majority of the user base of most online communities². Morzy investigates micro-communities, and finds that the forums clusters can be categorised into groups such as expert forums and socially coherent forums (Morzy 2010). Adamic et al. use a rudimentary clustering based on the three features thread length, content length, and asker/replier overlap, as a part of several analyses of user behaviour in Yahoo! Answers (Adamic et al. 2008). They conclude that certain types of topics cause specific user behaviour. For example, factual questions may cause short threads with few answers, whereas controversial questions might spark many replies and hence may cause long threads. However, the authors do not further investigate their intuition by formally defining and measuring factuality or controversy of questions.

Topic modelling approaches capture the semantic essence of a topic, e.g. (Bogdanov et al. 2013; Rowe et al. 2013), but the semantics do not tell us why people interact in different ways. In our work, we study the effects that certain topics might have on user behaviour, independent of their semantic meaning. We define three latent topic properties, namely Accessibility, Sociability and Controversy, and measure their correlation with user behaviour clusters.

3 Clustering Stack Exchange Communities

We investigate 29 online communities from the Q&A platform of Stack Exchange (www.stackexchange.com). For a definition of *community*, we rely on Stack Exchange’s notion, where each site is a community whose participants discuss a specific topic, such as programming, physics, and cooking. We extract 47 user behaviour features that cover aspects of user activity, user interaction and content creation, for example average posts per user, average in-degree, and average answer length, respectively. We then standardise the features, and apply a principal component analysis to minimise bias of strongly correlated features. Similar to Morzy; Chan, Hayes, and Daly, we use an agglomerative hierarchical clustering algorithm, and we define similarity between communities as the Euclidean distance between feature vectors of user behaviour. We describe the clustering approach in detail in (Aumayr 2016).

The result of the hierarchical clustering is a dendrogram that shows the similarity between the communities. Communities that are connected near the leaves of the tree are very similar to each other, whereas communities that are connected near the root are most dissimilar. Examining the clusters in Figure 1, we find that some communities that appear intuitively similar indeed exhibit similar user behaviour, and are therefore connected near the leaves of the dendrogram. Examples that demonstrate this well are the Android and Apple communities, and the Unix/Linux and Ubuntu communities. Other communities are somewhat related, although it is not obvious that they exhibit similar user behaviour, like the Maths and Tex/LaTeX communities (we assume that people who want to use mathematical formulae in LaTeX documents might frequent both communities for help). There are also communities that do not seem to be apparently related, like Bicycles and Electrical Engineering, and it is interesting to see that bicycle enthusiasts show online behaviour that is similar to electrical engineers.

The very popular Stack Overflow community is in its own cluster, which is most dissimilar to the other community clusters. In Table 1, we see that its popularity manifests in a faster growth and greater activity than we find in the other community clusters. We also notice that both the Stack Overflow community as well as the Programmers community are among the oldest communities in the data. The other communities, especially in the clusters “Non-IT” and “Science and its application” have a younger average age. We assume that the user behaviour changes over time, from a community’s inception stage towards its maturity.

Although there are some outliers, the clustering algorithm has grouped together many communities that appear topically related: Most of the non-technical communities are in

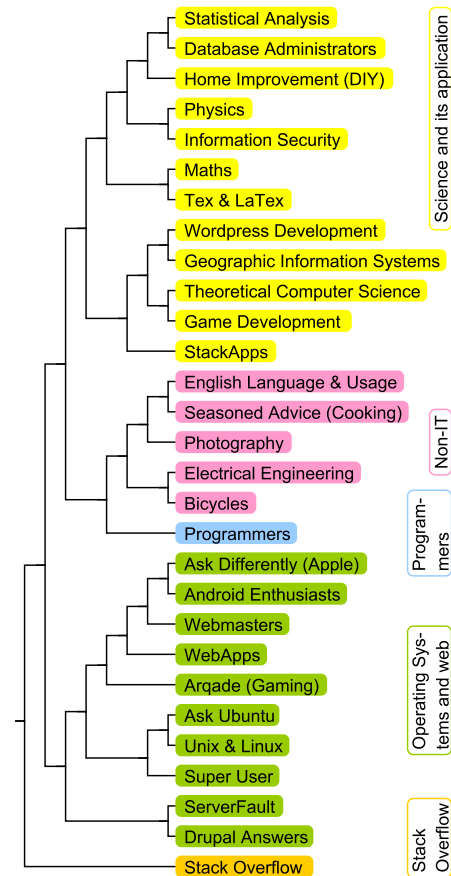


Figure 1: Cluster hierarchy of the 5 Stack Exchange clusters. We manually assign the cluster labels (vertical text).

their own cluster, and the remaining technical communities are divided into two categories that we manually label “Operating systems and Web” and “Science and its application”. It is especially noteworthy that many closely related topics are clustered together, although this is not a topic clustering.

4 Topic Properties and User Behaviour

We noticed that many of the communities in the same cluster also match topically, which suggests that there may be a relation between the topic and the way users interact, as also concluded by Adamic et al. In this section, we investigate the correlation between topic and behaviour, and thus, state the null hypothesis as follows:

H_0 : There is no statistically significant correlation between the type of topic and the community categories that are based on user behaviour.

Since we want to investigate whether there is a correlation to a satisfactory degree, our alternative hypothesis is therefore:

H_1 : The similar user behaviour in each community cluster is correlated to the type of topic.

We conduct two experiments to investigate this correlation. In the first experiment, under the assumption that

Cluster label	Properties	Access	Social	Controv.
Science and its application	Slow growth, low information spread	0.60 (M)	1.86 (M)	0.94 (M)
Non-IT	Slow growth, few ignored questions, answers longer than questions	0.75 (H)	2.02 (H)	1.07 (H)
Programmers	Many answers per question, few ignored questions, low connectedness	0.63 (M)	2.10 (H)	1.09 (H)
Operating systems and Web	Low information spread, low connectedness	0.53 (M)	1.72 (L)	0.95 (M)
Stack Overflow	Rapid growth, many Qs and As per user, few ignored questions, high information spread, low connectedness, As shorter than Qs	0.10 (L)	1.70 (L)	0.90 (L)

Table 1: Cluster properties and latent topic properties, indicating relative ratings: low (L), medium (M), or high (H).

people who talk about similar things also behave similarly, we use a bag of words approach on the message contents (punctuation and stop words removed). On these terms we compute the TF-IDF (term frequency and inverse document frequency) scores and then measure similarity between the communities via cosine similarity. However, a clustering of the communities based on term similarity has no resemblance with the clustering based on user behaviour, and we find that there is indeed very little overlap of terms between the communities. In conclusion, the bag of words approach is not suited to reject H_0 because the terms themselves are very specific for each community and have no correlation with the categories we found in Section 3.

In the second experiment we move away from the semantic plane and investigate the effects that different types of topics might have on the user behaviour. We therefore define the three latent topic properties:

- **Accessibility:** How difficult is it for new users to join the discussions based on the language that is being used?
- **Sociability:** How friendly and open are the people discussing the topic?
- **Controversy:** How frequently do original posts (e.g. questions) spark discussions and arguments?

Unfortunately, there is no widely accepted state of the art regarding robustly measuring these three topic properties in online communities, although there are some noteworthy works in the field of controversy detection (Whitworth and Felton 1999; Dori-Hacohen and Allan 2015). For the sake of reproducibility of our results, we experiment with a number of straightforward features for each latent topic property, and determine the best performing ones.

To measure Accessibility we look at the level of the **specialised language** that people use. Our assumption is that when they use a lot of words that are particular to the domain, then it requires more knowledge to take part in the discussions. We follow an approach described by Chung and Nation for each community: We determine the ratio of specialised vocabulary in a corpus (a community) by contrasting the occurrence of extracted words with their occurrence in a general purpose corpus (Chung and Nation 2004). Words that occurred at least 50 times more than in the general purpose corpus (or did not occur there at all) were marked as specialised words. We contrast the words from the Stack Exchange communities against the English Wikipedia as the general purpose corpus (obtained from dumps.wikimedia.org). It covers a whole spectrum of topics and domains, and fulfils the requirement of being at least 2

million words long, as specified by Chung and Nation. Last, we subtract the specialised language value from 1 so that high values represent a high accessibility.

As an alternative and simplified metric for specialised language, we also consider the **word repetition frequency**, which measures how often unique words are repeated on average. Another interpretation of this feature is the inverse of the number of unique words normalised by the total number of words. For all messages in each community, we extract and count the total number of words and divide it by the number of unique words. The rationale behind this feature is that the more words are repeated on average, the less time people need to grasp what is being talked about.

For measuring Sociability, we examine the tone of the discussions. Assuming that the friendliness and openness of people is reflected in a positive sentiment of their language, we use the sentiment analysis tool SentiStrength (Thelwall et al. 2010) to extract the positive and negative sentiment scores (each ranges from 1 to 5) from questions and answers. In particular, we are looking at the following four features describing Sociability: the **positive sentiment** and **sentiment difference** ($pos - |neg|$) in **answers** as well as in **all posts** (questions and answers).

For Controversy, we assume that a topic is controversial when people post heated arguments. Here, we employ SentiStrength to extract the **negative answer sentiment**, the **amplitude of both positive and negative answer sentiment** ($pos + |neg|$), and the **standard deviation of the answer sentiment difference** ($pos - |neg|$). The role of the standard deviation is to capture strong negative sentiment even when there is equally strong positive sentiment, in which case they would negate each other when only taking the sentiment difference into consideration.

From these nine latent topic property features in total, we select the strongest ones by calculating the correlation ratio η for each feature, which yields the ratio of a feature’s dispersion within a category divided by its dispersion across the entire sample set, i.e. the communities. This allows us to judge how well a feature is suited to represent the communities in their respective categories from Section 3. As a result, the strongest features are the specialised language for Accessibility ($\eta \approx 0.69$), the positive sentiment in answers for Sociability ($\eta \approx 0.71$), and the standard deviation of the answer sentiment differences for Controversy ($\eta \approx 0.67$). The limits of the correlation ratio are $0 \leq \eta \leq 1$, thus the resulting numbers between 0.67 and 0.71 indicate a moderate ability of our topic property features to associate the communities with their respective categories.

We perform a significance test on our hypothesis in order to prove that there is a correlation between the behaviour categories and the latent topic properties. For that, we transform the correlation task into a classification task, and devise a simple classifier that tells us how well the topic properties can reproduce the behaviour categories. First, we calculate the mean and standard deviation of each of the five behaviour categories regarding Accessibility, Sociability and Controversy. As an example, category “Non-IT” has a mean of 0.75 and a standard deviation of 0.08 for Accessibility (we record the mean values in Table 1). The classifier then predicts a category for a given community by selecting the category whose mean is closest to the community, under the condition that it falls within 3 standard deviations of it. Since there are three topic properties, the classifier repeats this three times for each of the 29 communities. If there is no agreement on a category between at least two of the three topic properties, the classifier selects one of them randomly.

If H_0 were true, i.e. there is no correlation, the classifier would not perform better than a random guesser with $p = 0.2$ for the five categories. We can reject H_0 if our topic property classifier is able to predict significantly more categories than a random guesser with an acceptable error margin of $\alpha = 0.01$. Our classifier correctly labels 17 of the 29 communities, which makes it unlikely that H_0 is true for $X = 17$ correct guesses at a random chance of $p = 0.2$:

$$\begin{aligned} P(\text{reject } H_0 | H_0 \text{ is valid}) &= P(X = 17 | p = 0.2) \\ &= \binom{29}{17} 0.2^{17} (1 - 0.2)^{29-17} \approx 4.67 * 10^{-6} \leq 0.01 \end{aligned}$$

In conclusion, by rejecting H_0 with a p-value $< 10^{-5}$, we can say beyond a reasonable doubt that the latent topic properties correlate with the behaviour categories. Table 1 lists the topic properties for each community cluster, and we can see that the clusters that are similar in user behaviour are also similar in latent topic properties (see also the similarity dendrogram in Figure 1). The Stack Overflow community stands out, with low topic property values in general, and extremely low Accessibility in particular. This is surprising, given that fact that it is by far the most popular community. However, The relatively high correlation between the topic properties themselves and the fact that our simple classifier performs with an accuracy of 58.6% indicates room for future improvements in the definition of the topic properties.

5 Conclusions and Future Work

In this work, we investigate the correlation between topic and user behaviour in online communities. We first cluster 29 Stack Exchange communities based on their similarity of user behaviour, and find that they can be clustered into five categories. These groups loosely represent intuitive, topically organised categories, which suggests that users who talk about similar topics also behave similarly. In order to capture the effects of different types of topics on user behaviour, we introduce the latent topic properties Accessibility, Sociability and Controversy. Then, we show that, although a bag of words approach provides no indication of different user behaviour, there is a correlation between latent

topic properties and user behaviour. Since we used simple ad hoc metrics to represent the latent topic properties, our goal for the future is to develop more robust and accurate metrics.

6 Acknowledgement

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

References

- Adamic, L. A.; Zhang, J.; Bakshy, E.; and Ackerman, M. S. 2008. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, 665–674. ACM.
- Aumayr, E. 2016. Categorising the online communities of stack exchange using quantitative user behaviour features. Technical report, National University of Ireland, Galway. <http://dx.doi.org/10.13025/S87P42>.
- Bogdanov, P.; Busch, M.; Moehlis, J.; Singh, A. K.; and Szymanski, B. K. 2013. The social media genome: Modeling individual topic-specific behavior in social media. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 236–242. IEEE.
- Chan, J.; Hayes, C.; and Daly, E. M. 2010. Decomposing discussion forums and boards using user roles. *Proceedings of the International Conference on Weblogs and Social Media* 10:215–218.
- Chung, T. M., and Nation, P. 2004. Identifying technical vocabulary. *System* 32(2):251–263.
- Dori-Hacohen, S., and Allan, J. 2015. Automated controversy detection on the web. In *Advances in Information Retrieval*. Springer. 423–434.
- Lazar, J., and Preece, J. 1998. Classification schema for online communities. *Americas Conference on Information Systems (AMCIS)*.
- Morzy, M. 2010. An analysis of communities in different types of online forums. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 341–345.
- Porter, C. E. 2004. A typology of virtual communities: A multi-disciplinary foundation for future research. *Journal of Computer-Mediated Communication* 10(1).
- Rowe, M.; Wagner, C.; Strohmaier, M.; and Alani, H. 2013. Measuring the topical specificity of online communities. In *The Semantic Web: Semantics and Big Data*. Springer Berlin Heidelberg. 472–486.
- Stanoevska-Slabeva, K. 2002. Toward a community-oriented design of internet platforms. *International Journal of Electronic Commerce* 6(3):71–95.
- Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; and Kappas, A. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61(12):2544–2558.
- Whitworth, B., and Felton, R. 1999. Measuring disagreement in groups facing limited-choice problems. *ACM SIGMIS Database* 30(3-4):22–33.