Analyzing behavior of posts belonging to different Subreddits

Siddharth Pathak

sidpath@iu.edu

May 4, 2018

1 Abstract:

In recent times, it has become a lot easier for people to get access to the Internet, and communicate with other people easily on broad topics. One of the means of communication is using social website like Reddit, where people post something, and have other people respond to the post, resulting in users having a threaded discussion. People can respond in different ways to different types of posts. To analyze this behavior, I built a network out of reddit posts from different subreddits, where each post and a comment is a node, and edges between the nodes which have a child and parent relationship between them. The size of the node is the number of upvotes the post has received. After building the network, I found that post on different subreddits differ from each other. Some posts attract more number of comments, which end up attracting more replies, thus having more focus on people having a conversation. While, on the other hand some posts' comments have more number of upvotes, than the post itself, signifying that the comments are of more importance than the post. Furthermore, some posts have less number of children, but a lot of upvotes. To quantify this result, I propose three `Key Parameter Indicators (KPI)`, each encapsulating different behavior: `Upvote Ratio`, `Average Depth` and `Average Number of Children`. Depending on these values, posts can be analyzed if they are more focused on discussion/dialog or are informative.

2 Introduction:

Usage of social media has increased exponential in the last few years. One of the famous social media platforms which has seen exponential rise is Reddit. On Reddit users post some text, links, images etc, and other users can comment and have a discussion in a threaded fashion. Reddit is further divided into topics, popularly known as Subreddits. Each subreddit is a portal to talk about certain topics. For example, few of the famous Subreddits are Soccer (people discuss everything related to Soccer) and AskReddit (people ask questions to Reddit users).

Prior work has been done to analyze the behavior of Reddit and its users. For example [1], studied how users of reddit are active on common subreddits i.e people who are active on "Soccer" subreddit are most probably active on `RedDevils` too (Nickname for Manchester United: a famous soccer club) subreddit. Further work [2] has been done to analyze discussions of a post and model them into Volume, Virality, and Responsiveness. A post has high volume means there are lot of active users, thus making the number of comments and posts high. Virality means if a comments had attracted a lot of attention, resulting in a huge thread. A subreddit is responsive, if the users reply to other posts or comments immediately in a short interval of time.

Different kinds of posts on different subreddits vary a lot from each other. Some of the posts have a lot of upvotes, but less comments i.e people tend to discuss less on such posts. While, some posts have a large number of comments and lot of threads in those comments. Furthermore, few posts have the number of upvotes in the same range as those of their comments, signifying that replies to the post are as important as the post itself.

Mapping these characteristics of different kinds of posts results in different types of networks. In the network, each post and comment is a node, and edges are present between nodes which have a child and parent relationship between them. The size of the node is the number of upvotes the post or comment has received. Posts which are more 'Knowledge' based or 'Informative' tend to have a lot of upvotes, but less number of comments. On the other hand, if the post is something which is asking a question to the users, then it will tend to have a lot of comments, and these comments tend to have number of upvotes in the same range as of the original post.

3 Methods:

3.1 Dataset:

Reddit provides data of each post directly in `JSON` format. Loading a Reddit webpage and appending the URL with `.json` extension gives out `JSON` data instead of HTML. I downloaded this data for posts belonging to different subreddits, and parsed it using Python. The data contained several different attributes, though I used only the number of upvotes, the poster, post content, etc. I also calculated the maximum depth for each post/comment, and the total number of children while parsing the data recursively. Once calculated, I created a `GML` file for reading the data into GEPHI for visualization. While creating `GML` file, I set the following attributes per node: `Score`, `Total Children`, `Maximum Depth`, and `User`.

3.2 Graph Analysis:

The structure of graphs varies from subreddit to subreddit. Posts on some subreddits attract more discussion from users, which ends up having more number of comments and deeper threads. These types of networks have high depth, and large number of children. These kinds of posts have comments which attract more user traffic, than the post itself. On the other hand, some posts have few comments, and these comments have less number of upvotes as compared to the post itself. To measure this different kind of behavior and to quantify it, I calculate three different types of `Key Parameter Indicator: Upvote Ratio, Average Depth` and `Average Number of Children`.

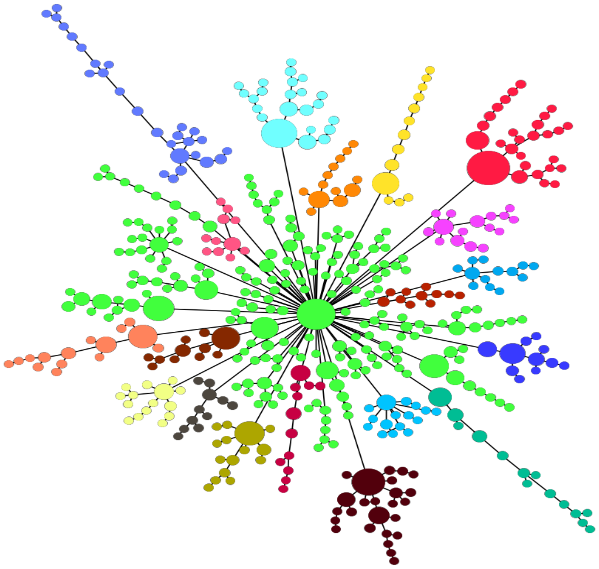

Figure 1: Graph of a AskReddit Post

The above figure is a graph of a typical post on AskReddit. The root node has high number of upvotes, while also few of the comments have high number of upvotes. The comments which have high number of upvotes also have high number of children, and high depth. The posts on AskReddit are discussion based, as the post itself asks questions to the users. The answers themselves attract user traffic in terms of upvotes and replies. This results in a graph having high depth, branching and several big size nodes as seen in the above figure.

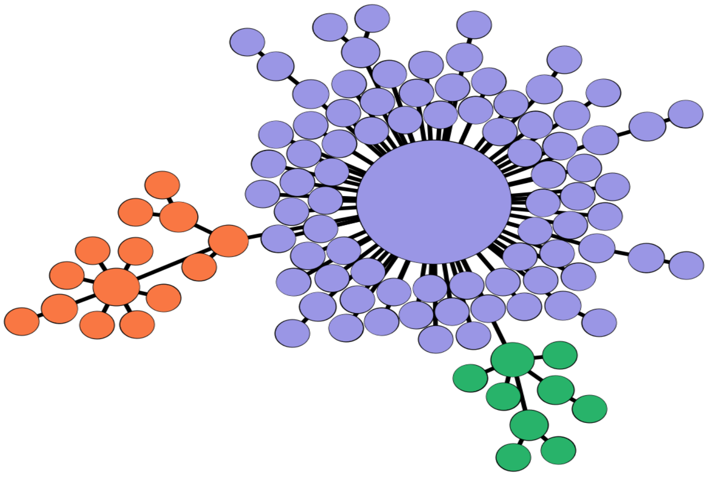

Figure 2: Graph of a LifeProTips Post

The above figure is a graph of a typical post on LifeProTip. The root node has high number of upvotes which is few dimensions higher than the upvotes of its children. The children have less number of upvotes, children and depth. This is because the posts on LifeProTips don't incite users to start a discussion or reply to the post, because the post itself is informative. Users read it, and agree/disagree with it. This results in graph having one huge size node i.e root, and all children of small size.

To quantify these results, I calculate three `Key Parameter Indicators` as follows:

- `Upvote Ratio:` It is the ratio of average number of upvotes of the top 10 percentile of comments based on their scores, and the number of upvotes of the post. If the number has less value, then the post is more important in terms of attracting traffic than the comments.
- `Average Depth:` Depth is the maximum length of a thread. Average depth value is the average depth of the top 10 percentile of comments based on the number of upvotes. If the average depth of the post is high, then the post is more discussion based and has more dialog/conversation.
- `Average Number of Children:` It is the average of total number of replies of the top 10 percentile of comments based on the number of upvotes. If there are more number of children, then the comment has bought in a lot of traffic i.e. a lot of users have replied to the comments of the post.

4 Results and Discussion:

I analyzed posts of five different subreddits.

Subreddit Name	Upvote Ratio	Average Depth	Number of Children
AskReddit	0.38	4.1	9.42
Soccer	0.06	3.47	11.92
LifeProTips	0.0005	0.49	1.31
EarthPorn	0.004	2.33	3.8
AskMeAnything	0.11	4.44	14.64

Table 1: Results for different subreddits

Based on the networks formed and the results in the above table, two of the subreddits (LifeProTips and EarthPorn) focus less on dialog and discussion between users, and more on the post itself. These kinds of posts are more informative, and don't incite the user to start a discussion or thread. LifeProTips has posts about life hacks and tips, which makes the post informative. The user either agrees with the tip, and upvotes it, or downvotes it. Similary, posts on EarthPorn are about landscape photography. If users like it then they upvote it, or downvote it. Thus, there is less probability of comments, and discussion. As seen in the results table above, Upvote Ratio for posts belonging to these subreddits is very less (couple of orders of magnitude less than the others) : 0.0005 and 0.004 as compared to the upvote ratio of AskReddit, Soccer and AskMeAnything, thus indicating that the number of upvotes posts get is very high than the average comments upvote. Similarly, Average depth and Average number of children is less, 1.31 and 3.8 respectively. Posts belonging to EarthPorn have slightly more number of children and higher average depth than LifeProTips, signifying there is some discussion, but not as much as other discussion based posts.

The other three subreddits which I analyzed are Soccer, AskReddit and AskMeAnything. Based on the networks formed and the above results, these subreddits have high depth, and large number of children i.e they focus more on discussion and dialog between users. They have high values for Average Depth, and number of children as compared to informative subreddits like LifeProTips and EarthPorn. Analyzing the results show that post on AskMeAnything have the highest average depth (4.44), and number of children(14.64). This is because the author of the post is a famous personality, whom the users ask a question, and he/she answers them, thus starting a discussion thread on the that answer. Similarly, AskReddit's post have high average depth and number of children, because the post itself asks the users of reddit some questions, and the answers spark a discussion and conversation between the users. Posts on Soccer are about Soccer matches, and it is more focused on discussion, as people discuss about how the teams tactics and performance. The upvote ratio for the posts belonging to AskReddit and AskMeAnything is high (0.38 and 0.11 respectively) as compared to the other informative posts, indicating that the comments and discussion which ensued because of the post are also important as the post itself. Upvote Ratio for Soccer is less than AskReddit, because the discussion is among people who are fans of certain teams, and users downvote comments which praise their rival teams.

5 Conclusion

The results show distinct difference in values of `Upvote Ratio`, `Average Depth`, and `Average Number of Children`. Depending on the range of the values, we can classify whether a post is discussion based, or informative. If the value of `Upvote Ratio` is very less(in the order of 10^{-3}), then the post has a lot more number of upvotes than the comments. If the value of Average Depth and `Average Number of Children` is high (greater than 8) then the post is discussion based, where users are in a dialog and there are several conversation threads going on.

Future work can involve analyzing different types of subreddit. Few subreddits might be discussion based, but also have very low `Upvote Ratio`, indicating they are also informative. Further analysis of different subreddits can help understand these types of subreddits.

One of the limitations is the number of comments I have used per post. I used the top 500 comments as sorted by reddit itself. Using more data can provide further insights.

References

[1] Randal S. Olson, Zachary P. Neal. Navigating the massive world of reddit: using backbone networks to map user interests in social media

[2] Daejin Choi, Jinyoung Han, Taejoong Chung, Yong-Yoel Ahn, Byung-Gon Chun, Ted "Taekyoung" Kwon. Characterizing Conversation Patterns in Reddit: From the Perspectives of Content Properties and User Participation Behaviors

[3] Visualizing discussions on Reddit with a D3 network and Embedly

[4] Mapping reddit's active communities. David Marx