Reddit comments analysis

Carel Kuusk*
*University of Tartu

***Abstract*—Online social networks have become hugely popular in this century, with Reddit being one of the largest. Reddit is a community aggregator, where people can join their own communities called 'subreddits'. Reddit also provides an easily accessible API and access to the data from the beginning of the site. This makes it possible to analyse time-evolution of early communities using a Reddit comment dataset scraped via the public API. The evolution of the network is characterized. The code and data is available on Github.**

***Index Terms*—Social network analysis, Reddit, network evolution**

I. Introduction

With the advent of internet, social networks started to gain popularity around the turn of the century. The first major social network was SixDegrees.com, which launched in 1997. However, the rapid growth came with wider adoption of the internet and mobile devices, with Facebook being launched in 2004, Reddit in 2005 and Twitter in 2006 being one of the most popular ones still active today. [1]

Reddit is a massive social network, which ranks amongst the top 20 visited sites on the web by traffic. Reddit is organized in user-created communities called subreddits. Each subreddit has its own set of rules in addition to the global moderation policy of the Reddit website itself, a common topic and a set of volunteer moderators who are allowed to enforce the rules. Users can subscribe to subreddits, in which case the activity on the subreddit appears in their personal feed. Each user can post on the subscribed subreddit, and comment on posts and other comments, thus engagement under a post follows a tree structure. Users can upvote and downvote posts and comments, the difference between posts and comments results in the total score for the post.

However, the early Reddit was a totally different website, being more similar to HackerNews in its organization. Users were not allowed to create their own subreddits before the start of 2008 and at first there was just one 'subreddit' called *reddit.com*. It was not a subreddit in contemporary sense, but instead just the front page where everybody could post and comment. The first subreddits were created manually by the site administrators, with the oldest proper subreddit being r/features or r/nsfw created on XXX.

However, the first subreddits, that acquired significant engagement, were r/programming (created XXX) and r/science (created XXX). This indicates the first audience Reddit was directed towards were predominantly male and nerdy. The early culture probably helps to explain why to this day Reddit's audience is mainly comprised of young males [2].

Reddit is relatively open to data acquisition compared to other large social media platforms like Facebook and Twitter. Reddit has a comprehensive API that anybody can use to acquire data from the very inception of the Reddit platform. Queryable entities include subreddits, users, links and comments, etc [3]. This project used a comments dataset scraped using the publicly accessible API.

The availability of easily accessible data even from the very beginning of a social media platform is also the main reason for choosing this project. The easily accessible and well-formatted data allows us to analyze how an early social networking cite evolved from a small site to the behemoth it is today. The initial plan was to analyze the evolution of different subreddits – how users discover subreddits, why some subreddits become more popular, interaction between different subreddits etc.

However, during the analysis several hindering factors were discovered. First, the original Reddit did not have any user-created subreddits, and only opened the platform for users to create their own subreddits in 2008. This subsequently resulted in exponential growth in both the number of subreddits and the number of users (as opposed to the previous roughly linear growth). In and of itself it is an interesting result and confirmation of the power of user-generated content that is so widespread in todays online activities. Sadly, the exponential increase in the number of subreddits and users also meant that the computational capabilities accessible to the author were not sufficient to analyze such vast volumes of data.

II. Related work

Possible analysis methods are provided in Cordeiro et al. (2018): snapshot analysis to identify more static features of the network structure and aggregate analysis (either via landmark windows or sliding windows) for analysing statistical properties of the graphs within a certain timeframe.

Troy Steinbauer has analyzed the overall properties of the Reddit network in 2011. He identified that for example related subreddit distribution was following the expected power-law distribution.

The behaviour of individual users was analyzed by Thukral et al in 2018. The analysis was partly done on the 2008 and partly on the 2014-2015 period. They discovered, that based on the commenting and posting patterns, the users can be clustered into three types based on when most of their contributions are made (either within a short period from sign-up, stably, or more actively after a long hibernation). Also they discovered a clear separation between commenters and posters.

III. Dataset

The dataset consists of all of Reddit's comments from 2005 to 2008 (included), that is from the inception of the Reddit.

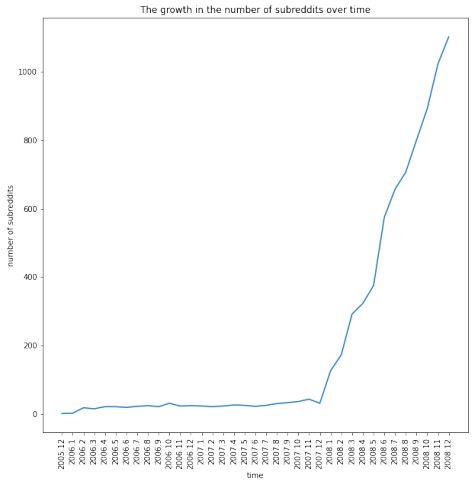
This provides an incredible opportunity to investigate initial stages of an online social network. The data was scraped by Reddit user u/Stuck_In_the_Matrix in 2016 and torrented from the Academic Torrents website. The original dataset includes comments up to at least 2015, but it would not have been feasible to include the whole of Reddit's comment dataset. However, the compressed dataset size is still 932 Mb, and covers the initial growth phase of the site, including the first year during which Reddit let users create their own subreddits.

Each line in the decompressed dataset contains data about one comment and is organized as a JSON document. The relevant fields are as follows:

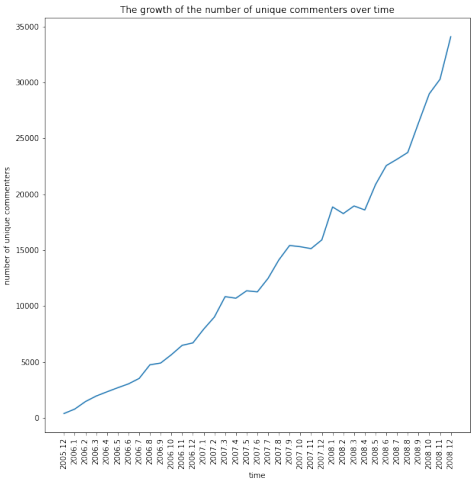
- `subreddit` – the subreddit under which the original link is;
- `subreddit_id` – the ID of the subreddit;
- `author` – the username of the author;
- `body` – the body of the comment;
- `score` – the karma on the comment;
- `link_id` – the ID of the link under which the comment was posted;
- `parent_id` – the ID of the parent, which can be either the link or another comment;
- `id` – the ID of the comment itself;
- `created_utc` – the UNIX timestamp of the comment creation.

There are more metadata, e.g. a field `controversiality`, which measures whether a comment has received a similar number of upvotes and downvotes, etc. Also, during the first years there were less fields (downvotes for example did not appear before 2008), so the selection was partly motivated by ensuring that all the fields would be present during the whole period under analysis.

Some basic cleaning was applied to the data. First, the fields `link_id`, `parent_id` needed some preprocessing to be compatible and comparable with the field `id`. Secondly, due to a relatively large number of comments authored by now-deleted users, some preliminary analysis steps required eliminating comments that were authored by the deleted users. This is because the name of the deleted comments in the Reddit's database is `"[deleted]"`, i.e. the statistics about user comment distributions, max number of posts, etc were skewed due to the there being a lot of comments authored by a deleted user (approximately 30% of the comments).



From the figure above we can see that the number of subreddits grew rapidly and then exploded at the start of 2008 due to Reddit allowing users to create their own subreddits. However, in subsequent analysis only the data up to March 2007 were used due to the volume of data.



The steady but fast growth of users is also clear from the data.

IV. Methodology

The methodology in this project follows the steps below.

1) Finding and downloading data.
2) Preliminary preprocessing, ensuring the high quality of data.
3) Preliminary data analysis, investigating potentially interesting further research paths.
4) Create a co-commentator graph.
5) Analyze time-evolution of network properties.
6) Analyze the possibility of applying a link prediction model on the graph.

Co-commentator graph is formed as follows. Each comment contains a `link_id`, i.e. the post ID under which the post

was made. All the comments under each post were collected and the authors of the comments were extracted. Then, each author was added to the graph as a node, and an edge was formed with all the other users who commented under the same post, i.e. with all the other co-commentators.

Depending on the analysis, this was done separately for posts under a specific subreddit or for a specific time-period (generally a month).

V. Results

The most interesting results were related to the first popular non-frontpage subreddits – science and programming. It appeared that people who commented under those subreddits had much higher centrality scores (eigenvector and PageRank), and kept commenting month-over-month. This is significant, since the comparative dataset did not include all users, but users who had already commented under a post with at least one other comment, i.e. the users who were not interested in the website and just signed up were excluded already. So the subreddits were clearly integral to the success in Reddit by drawing more engaging users and interactions.

VI. Conclusion

References

[1] B. Howard, "Analyzing online social networks," *Commun. ACM*, vol. 51, no. 11, p. 14–16, Nov. 2008. [Online]. Available: https://doi.org/10.1145/1400214.1400220

[2] M. Duggan, "6% of online adults are reddit users," *Pew Internet Am Life Project*, pp. 1–10, 2013.

[3] "Reddit api documentation," https://www.reddit.com/dev/api, accessed: 2021-06-01.

Note: For the Summary paper submission only, references to the authors own work should be cited as if done by others to enable a double-blind review. **Citations must be complete and not redacted, allowing the reviewers to confirm that prior art has been properly identified and acknowledged.**