Visualizing Reddit: Exposing user communication patterns through data visualization on Reddit.com

Jessica Peter

user behavior, visual analytics, online communities, reddit

Participation in online communities has become a significant part of many people's lives due to the popularity of forums, social media, and other types of websites and applications. With this participation comes a wealth of data in the form of textual information and metadata. This data could be useful to a variety of parties: From community members interested in better understanding the dynamics of their favourite online hang–outs, to moderators wishing to identify potential problem members, to marketers looking to identify the best times of day and verbal phrasing to appeal to a particular group of online users. However, the abundance of data can act as a double–edged sword: With so much data generated every day, users may have difficulty identifying patterns in user communication trends. In this paper I present Visualizing Reddit: a visualization–based interface I developed to help users identify communication patterns in online communities based out of Reddit.com. I discuss the unique capabilities of Visualizing Reddit for accessing a variety of metadata from a six–month timespan, as well as the design decisions that informed my development.

1 Introdução

For many, online communities hold the same personal significance as geographical ones, but without visual feedback, and amid the participation of sometimes millions of other users, the significance of each user, and his or her relationship to the whole, can become difficult to conceptualize. I was intrigued by the potential of data visualization as a means of solving this problem. Data visualization makes use of the computational capacity of computers to collect and sort huge quantities of data, but the production of insight is still up to the viewer, potentially facilitating a variety of tasks.

I chose to visualize online communities based out of the social news website Reddit.com for a number of reasons. For one, Reddit hosts a number of types of online communities, known as "subreddits." This gives me the opportunity to study a wide variety of communities. I was also interested in incorporating Reddit's "karma" feature into the visualization. When users submit original posts or comments, other users can vote on that content by clicking on either an upwards–pointing arrow or a downwards–pointing arrow to indicate if they like or dislike the content. The content's score, calculated based on these votes, is called karma. Analyzing karma could provide insight into whether content was well or poorly received by other users.

2 Visualization

Choice of Subreddits

When selecting subreddits, I wanted to represent a diverse range of the types of online communities someone might interact with. Based off of categorizations proposed by Ridings and Gefen (2004), and Lazar and Preece (1998), as well as my own observations, I decided on three dominant (though not exhaustive) categories of online communities to study: skill–building, fandom–focused and location–based categories. I selected three subreddits that fell into each of these three categories. This selection consisted of portal, scrubs, and tolkienfans for fan–focused; linux4noobs, sewing and learnart for skill–building; and vancouver, washingtondc, and thenetherlands for location–based. All of these subreddits had approximately 30000 subscribers at the time of selection.

User Social Roles

A recurring element throughout Visualizing Reddit is the grouping of users by their probable social role in the subreddit. This aspect of the visualization was inspired by the work of Viégas and Smith (2004), and by Smith and Wesler (2006) on identifying users' social roles on Usenet. I derived the social role categories from Turner et al. (2005), though I adapted the categories to be more compatible with the language of Reddit. The social role categories I came up with are as follows: submitters, commenters, trolls, conversers, and casual users. I introduced the category of casual users since I found that a large number of users only contributed a few pieces of content within my six month timeframe. The functional definitions I used to predict likely social roles are seen in Figure 16.

Table 1: Functional definitions of user social roles used in Visualizing Reddit.

| Social role	Definition
Casual	All users who have contributed three or less pieces of content
Troll	Users who have an average karma of less than –1
Submitter	More than 75% of content contributed is submissions
Commenter	More than 75% of content contributed is comments
Converser	Between 25% and 75% of content contributed is submissions

Note that each user may only belong to a single social role group: if a user is categorized as a casual user, s/he cannot be a troll; if a user is categorized as a troll, s/he cannot be a commenter.

Each user role is calculated using the entire six month dataset since I deemed that it would be useless to calculate a user's likely social role based on only the selected timespan, particularly if the viewer selects a timespan of a day or two.

Visualizations

Two different visualization modes make up Visualizing Reddit. Though these modes have no official names, I will be referring to these modes as "User Mode" and as "Thread Mode." The visualizations were made using Data Driven Documents (d3).

Figure 1: Screenshot of User Mode displaying all subreddits.
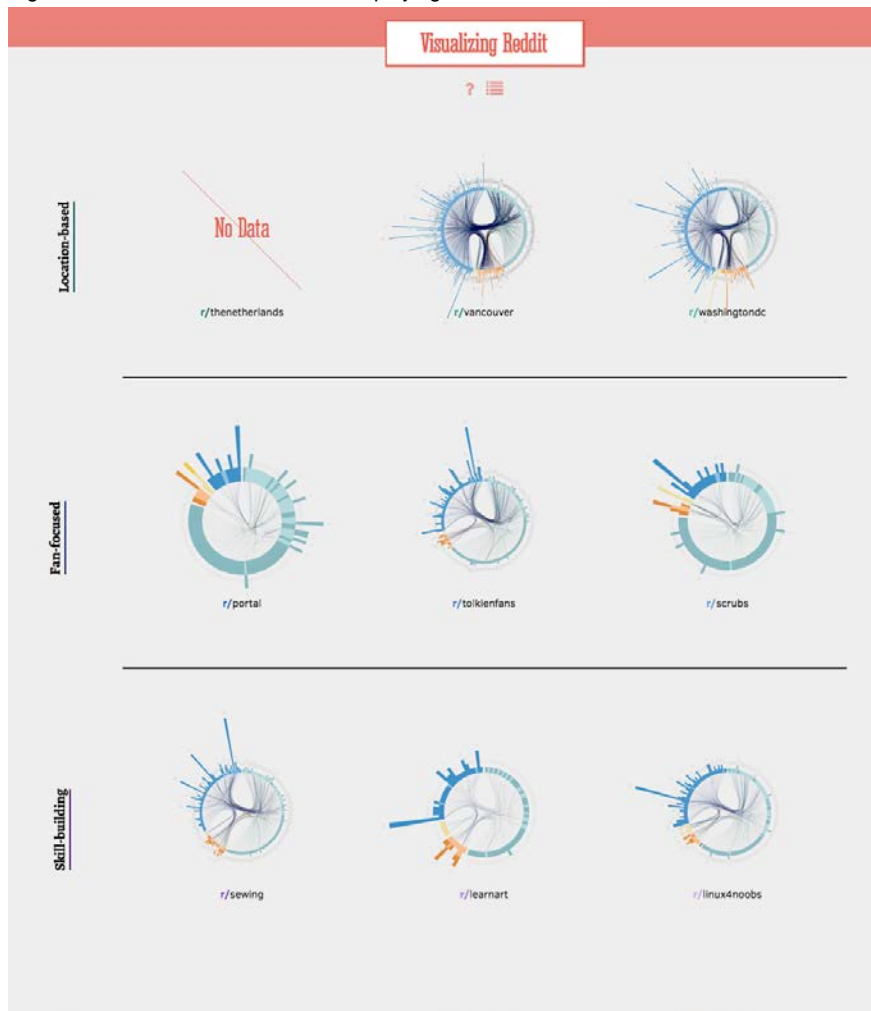
User Mode

In the first mode, subreddits are visualized using a hybrid chart composed of polar area–type charts and lines indicating user interaction, grouped using a hierarchical edge bundling algorithm (see Figure 1). The users are grouped according to their calculated social role, with the colour of the arc acting as the label. The lines connecting one user arc to another are coloured according to the user who created the interaction (e.g. the line would be green if a green casual user commented on a yellow submitter's post).

Figure 2: Diagram explaining how to read a stacked polar area chart.
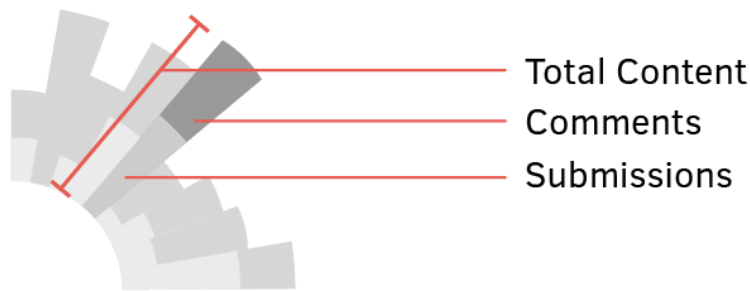
Each arc is further divided according to the type of content the author contributed. Submissions (as opposed to comments) are a slightly desaturated version of the social role colour (see Figure 2). On the outside of each arc is a number indicating the total number of pieces of content the author contributed.

I chose this chart because it allows the viewer to mentally divide the data into categories based on user social role, user, and the user's type of contribution (submission or comment). The lines at the centres of the charts may help viewers identify if specific individuals or groups of users interact regularly, as well as users who tend to receive a lot of comments from other users.

Figure 3: Screenshot of Thread Mode.
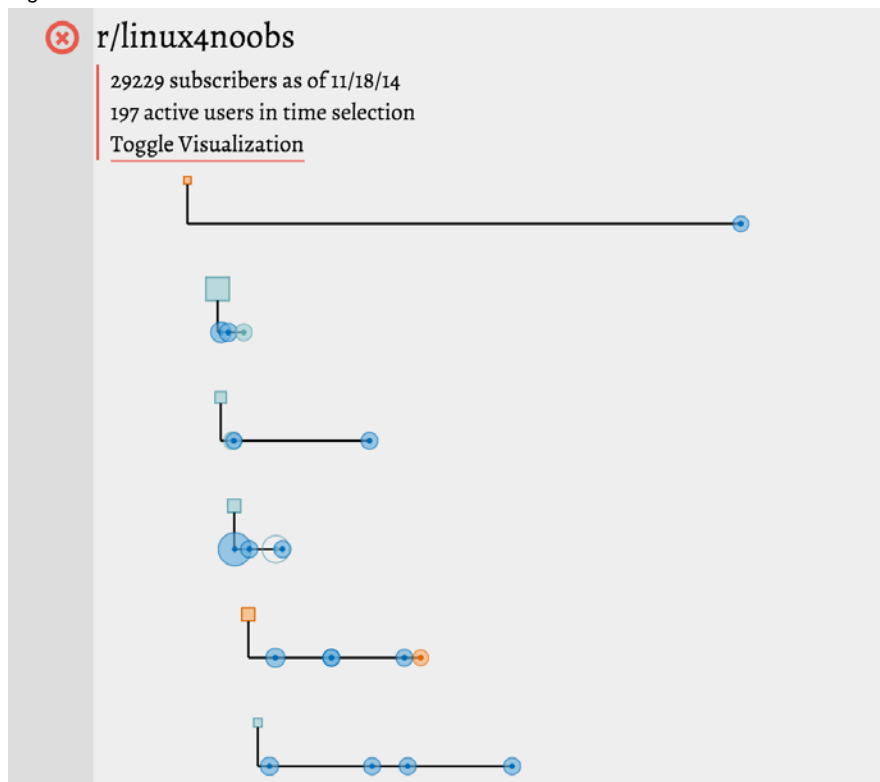
Thread Mode

Thread Mode can only be accessed when the viewer is examining an individual subreddit. To access this mode, viewers click a link labeled "Toggle Visualization" in the top right of the page. In this mode, an L-shaped line represents each individual thread. The length of the line represents the time between the initial submission and the last comment made to that

submission scaled according to the user-selected time frame (see Figure 4; see "Interface design and interaction" for a discussion of time range selection). On the left side of the subreddit, there is a square, representing the original post. Circles appear along the length of the line. Each circle represents a comment. The size of the shape represents the absolute karma value of the submission or comment. If the karma score of the content is greater than zero, the shape is a solid colour. If the karma score is zero or less, the shape is represented by an outline. As in User Mode, the shapes are coloured according to the user's social role, though there is no variance in saturation.

Figure 4: Diagram explaining how to read Thread Mode.
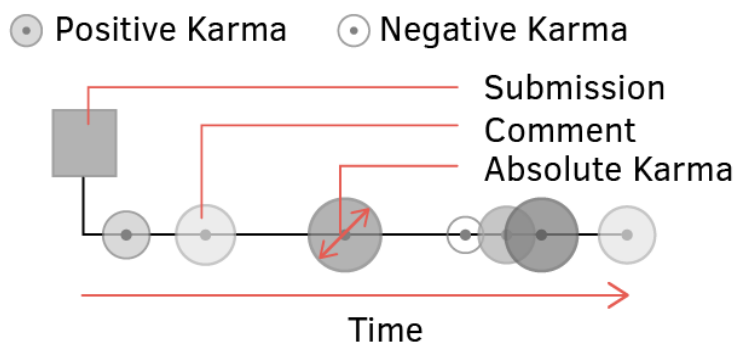
I designed Thread Mode so viewers can follow conversations in threads, and to enable the formation of hypotheses about threads, rather than about users or user groups. The viewer can make conjectures on, for example, what kinds of posts receive the highest or lowest karma in each subreddit, or which kinds of submissions receive the most comments.

**Interface design and interaction**

On initial load, users see nine different charts organized into rows. Users can click to bring one of the individual charts to a full–screen view. Here, the user can click on arcs representing users to view more information on each user. This interaction causes the selected arc to become opaque and the lines showing whom that individual user has interacted with to turn red. A pop–up appears to the right of the visualization containing information on the selected user (see Figure 5). This pop-up includes a directional node–link diagram showing all the users the selected user communicated with during the time frame. If the user clicks on a shape in Thread Mode, a similar pop-up appears featuring the post's content, the author's name, and the karma score.

Figure 5: When the user clicks on an arc, a pop-up with information on the selected user appears next to the chart.
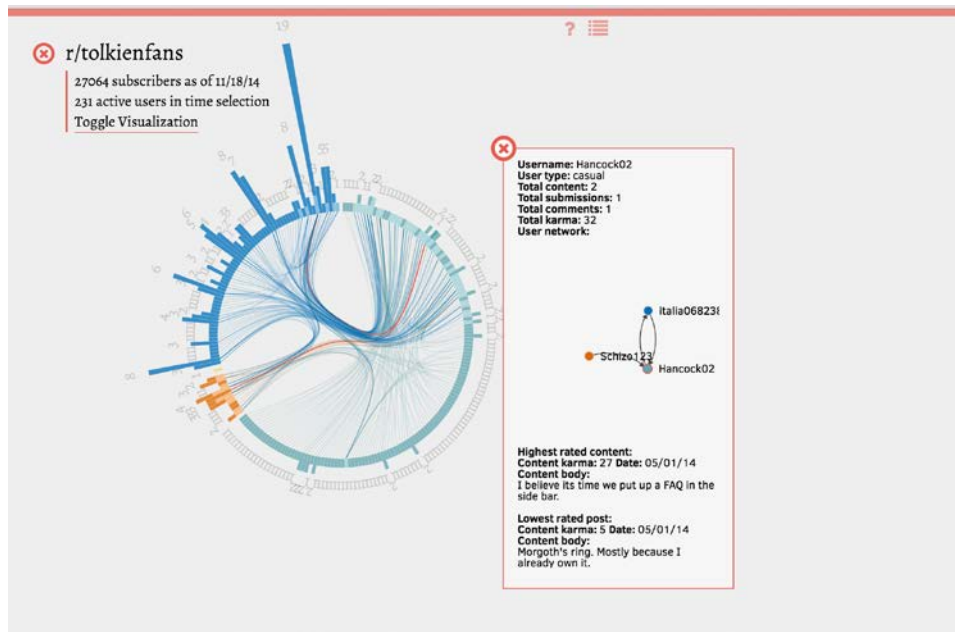
Initially, the charts reflect the first week and a half of the six–month dataset. Users can access the full range of time by clicking the menu button at the upper right of the screen. The timeline appears at the bottom of the screen as a colour–coded stacked area chart with the different stacks of colour representing content contributed by a particular subreddit (see Figure 6). Users can change the visualized time range by dragging a semi–transparent rectangle across the timeline, or changing the rectangle's size.

Figure 6: The menu with interactive charts and timeline pen.
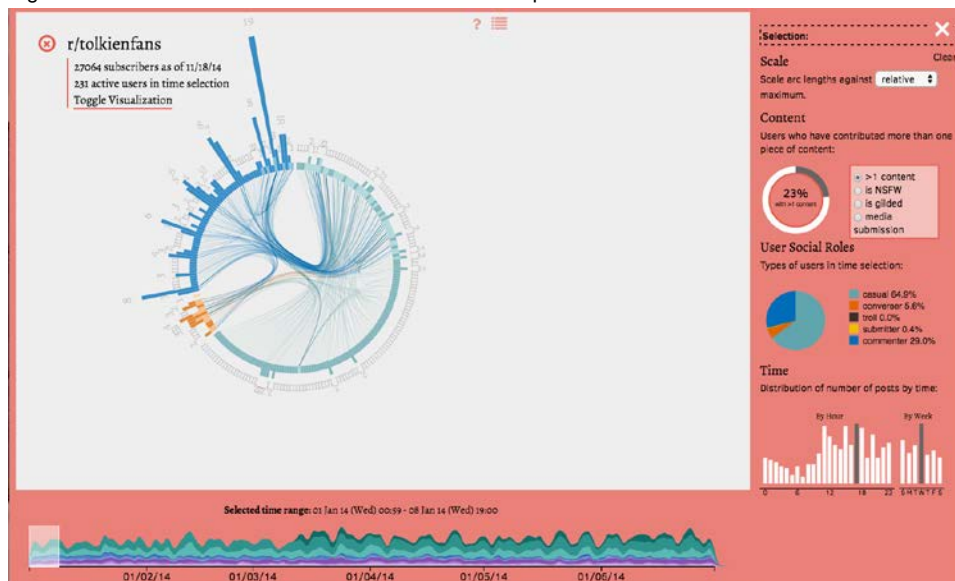
On the right side of the menu there is a range of charts that show statistics related to the time selection. If the viewer has isolated a single chart in full–screen view, the data shown in the charts reflects only the selected subreddit. If the viewer is looking at all nine subreddits, the information reflects user behaviour across all subreddits.

By clicking pieces of these charts, viewers highlight users who fit the selected criterion. For example, if the viewer clicks the slice of a pie chart representing users who have contributed more than one piece of content within the selected time frame, users who do not meet these criteria fade out and are unselectable.

Figure 7: The viewer has chosen to highlight only users who have contributed more than one piece of content.
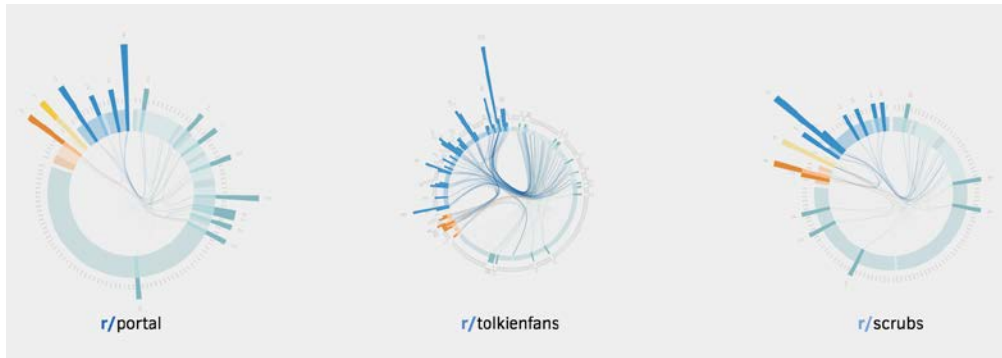
The ability to highlight users based on whether they fit selected criteria was so that the viewer can use the charts as filters for the data as well as use them as a summary report on features of the dataset.

3 Conclusions

Visualizing Reddit allows viewers to navigate online community archives through unique controls. Not only are viewers able to access the textual content of previous conversation threads, they can also use filters and controls related to the content's metadata to potentially reveal more complex patterns.

Though I have conducted a short pilot test with members of my research lab, I have not yet completed more rigorous testing to reveal how viewers use Visualizing Reddit. I would like to make it possible for users to add subreddits and categories of their choice. I would also like to enable real–time updating so that viewers can use Visualizing Reddit for moderation or "feed–reading" purposes.

References

FISHER, D., SMITH, M., & WELSER, H. T. 2006. *You are who you talk to: Detecting roles in usenet newsgroups.* System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference.

LAZAR, J. & PREECE, J. 1998. *Classification schema for online communities.* AMCIS 1998 Proceedings.

RIDINGS, C. M. & GEFEN, D. 2004. *Virtual community attraction: Why people hang out online.* Journal of Computer‐Mediated Communication. 10 (1).

TURNER, T. et al. 2005. *Picturing Usenet: Mapping Computer‐Mediated Interaction.* Journal of Computer‐Mediated Communication. 10 (4).

VIÉGAS, F. B., & SMITH, M. 2004. *Newsgroup crowds and authorlines: Visualizing the activity of individuals in conversational cyberspaces.* System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference

About the author

Jessica Peter, BFA, OCADU University, Canada <jp11jg@student.ocadu.ca>