University of North Carolina Charlotte

Wikipedia Vandalism Visualization

Project Report

By

Kanchinadam Teja

Anveshi Joshi

Aditya Jumde

Department of Computer Science

Submitted in partial fulfillment of the requirements for the course

Visualization & Visual Communication by Dr. Aidong Lu

We chose Wikipedia Vandalism dataset and used 4 methods to visualize the data:

1. Bar & Pie chart visualization: Bars represent months Jan 2013 to Dec 2013 with a pie chart representing number of benign and vandal users for each month.
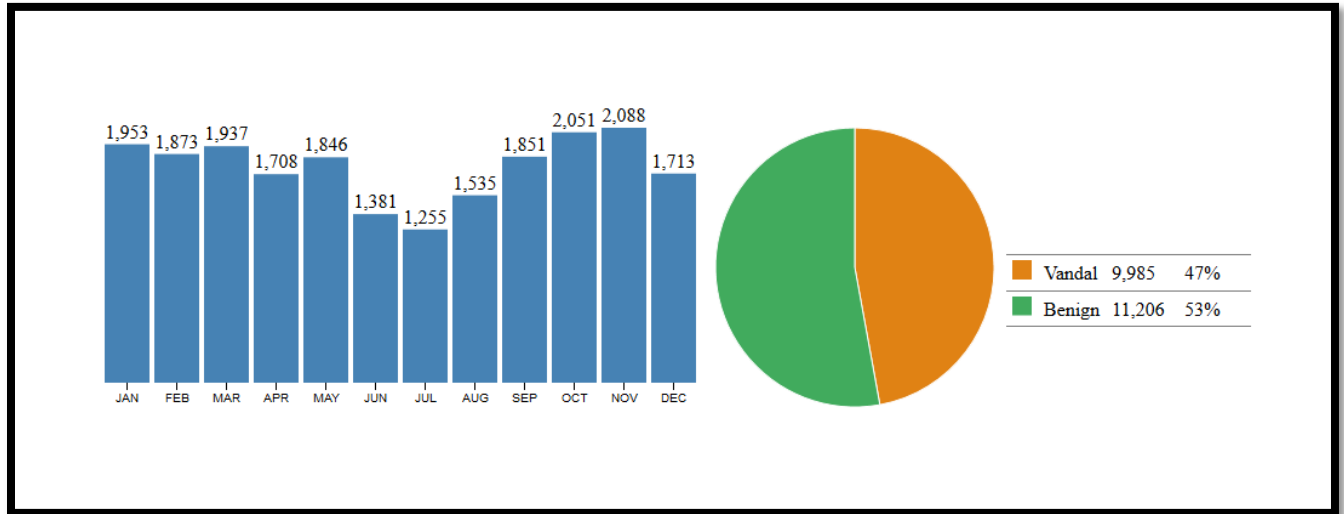
2. Cluster purity visualization: Each cluster represents three attributes of data namely bot score, reputation and intensity of users in each month. Intensity signifies the average vandalism in that particular month which is calculated by bot score which was mentioned in the database (refer the database readme file for details). Reputation is a score in database, it signifies average reputation of users in a particular month.
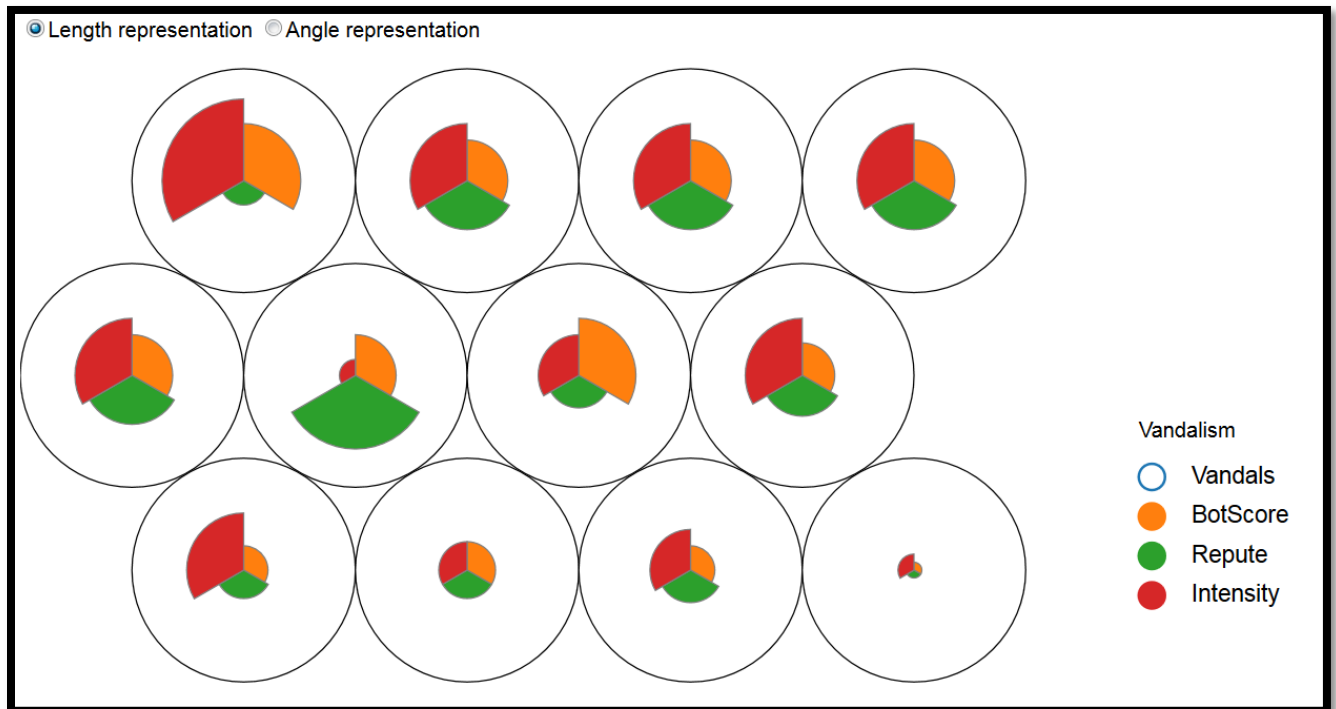
3. Bubble visualization: Each bubble size is proportional the number of times each web page was edited. Largest bubble signifies the web page being edited the most during that period of study and the smallest bubble signifies

the least edited web page in that particular period of time. When we hover over the bubble we get page name and number of times the page was edited.

Image1 has data about all web pages, and none of them is visible, so we decided to take top 100 webpage only.

Image 1

Image 2

4. Tree Visualization: Sorting is done in descending order based on sticki score. Each block has sub blocks displaying user reputation, average sticki score and the number of edits. When you hover over the block tooltip displays page title and average sticki score. This visualization is only for top 12 most vandalized web pages.