

Project Proposal

William Farmer
Edward Maton

November 8, 2013

1 Project Goals

The goals of this project are to create a service that allows users to analyze their Twitter profiles. These users will be able to create an account, log in to the site, and see very explicit details about their profiles, including how they rank compared to other users in terms of influence. For more details about data methods and algorithms utilized, please see A.

2 Views

1. Home Page
 - (a) Capabilities
 - i. Login Link
 - ii. Global Profiles Rankings
 - iii. Sample Visualizations
 - (b) Constraints
 - i. Only so much can fit on this page, and details will have to be known before we decide the exact content.
2. Login Page
 - (a) Capabilities
 - i. User/Pass Fields
 - ii. Login Button
3. Profile Details Page
 - (a) Capabilities
 - i. Show all user details. See Appendix A
 - (b) Constraints
 - i. Limited by Twitter API. Can only show certain details.
4. Network Visualizations
 - (a) Capabilities
 - i. Visualizations of social networks. Each node being a user, and line types indicate connection type.
 - (b) Constraints
 - i. Limited by D3 capabilities and network size. Too big and it will have to be limited.
5. Ranking Page
 - (a) Capabilities

- i. Show a list of all users queried on the site and how they rank next to each other. If less than 100 users have been queried, give sample users from dataset.
 - ii. Link to “Full” page. This ranks all users in the dataset.
- (b) Constraints
- i. Limited by data available.¹

3 Stakeholders

Those interested in the project would be on two fronts:

1. **Individual Users:** These users will be interested in how important they are compared to others, and how much they matter to the general public.
2. **Companies:** These profiles will be interested in how their message is spreading and how much users are talking about them. This is very important to many companies, as it’s a direct link to their advertising effectiveness.

4 Conclusion

This project will serve as a detailed Twitter profile information aggregator that also serves to rank users in terms of influence. The front end aspects of the project are fairly straightforward, and comparatively easy next to the algorithmic aspects.

We know the scope of the project, and are very confident that we can create a usable service that maintains effective functionality while being easy to use and reliable.

A Algorithmic Details

Our group wishes to analyze Twitter data and determine a ranking method in order to sort individual Twitter profiles based on influence. Each tweet looks like the following sample tweet:

¹Which isn’t a concern. We have a lot of it

```

1 {
2   "contributors":null,
3   "coordinates":null,
4   "created_at":"Fri Oct 18 02:21:23 +0000 2013",
5   "entities":{
6     "hashtags":[],
7     "symbols":[],
8     "urls":[],
9     "user_mentions":[]
10  },
11  "favorite_count":0,
12  "favorited":false,
13  "filter_level":"medium",
14  "geo":null,
15  "id":"391026191655661570",
16  "id_str":"391026191655661570",
17  "in_reply_to_screen_name":null,
18  "in_reply_to_status_id":null,
19  "in_reply_to_status_id_str":null,
20  "in_reply_to_user_id":null,
21  "in_reply_to_user_id_str":null,
22  "lang":"en",
23  "place":null,
24  "retweet_count":0,
25  "retweeted":false,
26  "source":"web",
27  "text":"<TWEET-TEXT>",
28  "truncated":false,
29  "user":{
30    "contributors_enabled":false,
31    "created_at":"Mon Feb 04 19:35:34 +0000 2013",
32    "default_profile":false,
33    "default_profile_image":false,
34    "description":"<USER_DESCRIPTION>",
35    "favourites_count":477,
36    "follow_request_sent":null,
37    "followers_count":1569,
38    "following":null,
39    "friends_count":1999,
40    "geo_enabled":true,
41    "id":"1148876598",
42    "id_str":"1148876598",
43    "is_translator":false,
44    "lang":"en",
45    "listed_count":6,
46    "location":"writer taken by Amon",
47    "name":"kitty pryde rp",
48    "notifications":null,
49    "profile_background_color":"642D8B",
50    "profile_background_image_url":"<URL>",
51    "profile_background_image_url_https":"<URL>",
52    "profile_background_tile":true,
53    "profile_banner_url":"<URL>",
54    "profile_image_url":"<URL>",
55    "profile_image_url_https":"<URL>",
56    "profile_link_color":"FF0000",
57    "profile_sidebar_border_color":"FFFFFF",
58    "profile_sidebar_fill_color":"7AC3EE",
59    "profile_text_color":"3D1957",
60    "profile_use_background_image":true,
61    "protected":false,
62    "screen_name":"wwefamily22",
63    "statuses_count":29869,
64    "time_zone":"Eastern Time (US & Canada)",
65    "url":null,
66    "utc_offset":-14400,
67    "verified":false
68  }
69 }

```

Where some fields have been abbreviated using < > in order to display correctly.

Each tweet holds a vast amount of information, and has all the ingredients needed to perform ranking analysis. We have scraped 1,977,401 tweets and 2,172,571 users using Twitter's development API, amassing to 1.2GB of data. This data was scraped in a 5 day period from the city of Boulder, giving a high percentage of related tweets.

Besides scraping tweets, we can also acquire profile information. This is much more limited due to Twitter's internal rate limiting system, however we could theoretically analyze a much smaller subset based on influence.

[illegible]

This network was created with a sample size of 10000 tweets. Each node that is visualized here is a retweet, and clusters appear when the individual that is retweeted is also in the sample. Conversely, the outer ring arises from the retweets that do not have a source tweet in the sample.

1. **user["id"]** - For uniquely identifying users.
2. **user["screen_name"]** - For uniquely identifying users in a readable format.
3. **user["followers"]** - Number of followers the user has.
4. **user["friends"]** - Number of friends the user has.
5. **text** - To identify retweets and possibly notable words/phrases.

CSCI 3308

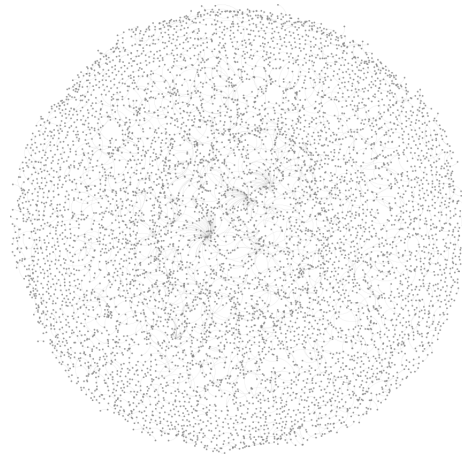


Figure 2: Gephi Visualization of 5000 retweets

friend acting as a “link”, or edge. This system relies on the eigenvalues of several matrices, and has high correlation with the material covered in this class.

The concept of ranking Twitter profiles on influence is interesting because Twitter holds a vast amount of information on our social lives, and by ranking individual users on their influence we can start to sift through Twitter feeds, identifying those that are the most important.

If we succeed in our endeavours, we will be able to not only sort through a set of twitter posts and rank the users on influence with respect to the others, we will also be able to take a given user profile and identify how influential the user is based on his twitter network and retweet percentages.

This project holds potential and can be adapted to a wide range of uses. It is also both accomplishable in the time allotted, and fits well into the scope of this course. We’ve been able to make enough preliminary progress as to determine that the dataset chosen is a good fit for this type of analysis. If completed well, we will be able to correctly rank users on their influence, a task that is not often thought of by Twitter users.