Finding your Doppelganger: Who's Your Most Similar Twitter Friend?

[COMP 440] Gozong Lor, Eva Yifan Gong, Tony Bach

Macalester College

Background

There have been many studies and apps on finding the similarity between any two users on Twitter. However, none give you the option to take in a set of friends of a user and find the one friend that is most similar to that user, so we wanted to create a program that can do this. Our approach of using profile similarity, network similarity, and content similarity is a combination of several approaches used in existing literature (Akcora 2011, Mizzaro 2015).

In addition, according to the theory of homophily, we tend to bond with people most similar to us, so we thought it would be interesting to look into whether people we interact with the most on Twitter are actually the ones we are most similar to.

Motivation

The widespread use of social media platforms like Twitter has increased the number of "friends" people can have, with Twitter users like @justinbieber totalling 71,116,651 followers and 245,597 friends. However, a majority of people's online interactions only involve a small subset of these followers and friends.

Our project and analysis will bridge that gap of having too many friends you don't know by providing users with a doppelganger who has similar speech patterns, profiles, and networks, but who may not be among the users you interact with the most.

Calculation Metrics

- Metrics:

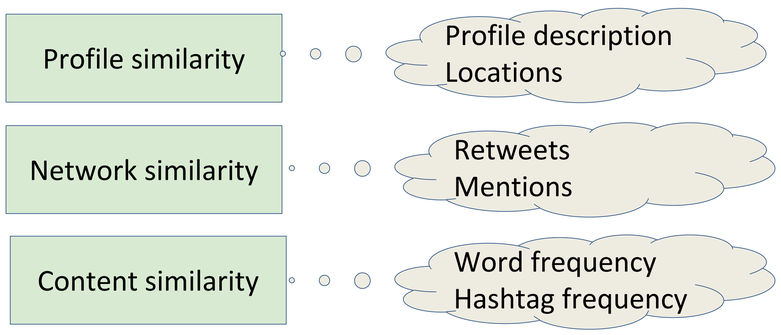

Figure 1. Metrics used to calculate user similarity

- Calculation Equation:

User similarity = x * Profile similarity + y * Network similarity + z * Content similarity

Data

We use Twitter's API to get data on a given user's friend list, mentions, retweets, profile information, as well as all their tweets and hashtags. However, due to the API rate limits, we can only get 180 friends and 200 tweets for each friend every 15 minutes.

References

Akcora, C. G., Carminati, B., & Ferrari, E. (2011, August). Network and profile based measures for user similarities on social networks. In Information Reuse and Integration (IRI), 2011 IEEE International Conference on (pp. 292-298). IEEE.

Mizzaro, S., Pavan, M., & Scagnetto, I. (2015). Content-Based Similarity of Twitter Users. In Advances in Information Retrieval (pp. 507-512). Springer International Publishing.

Profile Similarity

For every user, create a dictionary to store the profile description and location

↓

Calculate the cosine similarity between a given user's description and all other users' descriptions

↓

Calculate the location similarity: a score of 1 is given to users with same locations; 0 otherwise

↓

Profile Similarity=0.5*Description Similarity + 0.5* Location Similarity

User	A & B	B & C	A & C
Profile	~1.0	~0.2	~0.2
Location	1.0	0	0
Overall	1.0	~0.1	~0.1

Figure 2. Examples of Twitter profiles and the profile similarity score

Results and Conclusions

- The most similar friend of some interesting Twitter accounts, using the metrics above:

User	Doppelganger	Relationship to User
Donald Trump	Dan Scavino	Advisor
Hillary Clinton	John Buysse	Social Media Strategist
Emma Watson	Elizabeth Nyamayaro	Recruited user for HeForShe campaign
Cristiano Ronaldo	James Rodriguez	Teammate
Macalester College	Macalester Alumni	Alumni association

Note: x=0.2, y=0.4, z=0.4.

Figure 6. Most similar Twitter friend of some interesting accounts

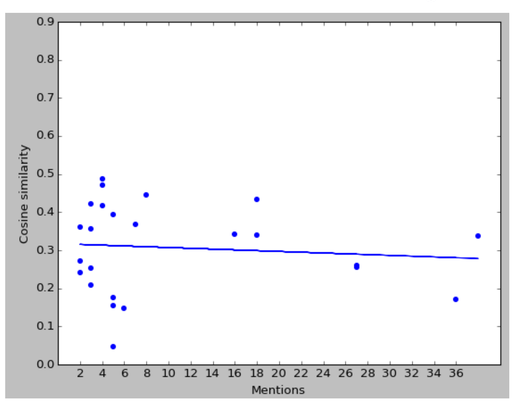

Figure 7. # of mentions and cosine similarity scatter plot

- We use the number of mentions as the independent variable , and the content similarity as the dependent variable to create the scatter plot above. As can be seen, there is no significant correlation between these two variables.
- The people most similar to you are not necessarily the people you talk to or interact with the most, and vice versa.
- You can use our app to identify the users you are most similar to that is not already in your immediate network, and get to know that person.

Content Similarity

- Tweet content similarity:

For every user, create a document of all tweets of that user

↓

Calculate tf-idf score of each word in each document

↓

Each document gets turned into a vector of tf-idf scores

↓

Calculate cosine similarity between each vector

Figure 3. Finding content similarity between users

- Hashtag content similarity (extra option): Same as above, but only use hashtags instead of whole tweets

Network Similarity

For a given user, find all the users they mention and retweet

↓

Generate a mention similarity vector for each mentioned user

↓

Calculate the cosine similarity for each user and the given user's mention vector

↓

Calculate cosine similarity between each vector

Figure 4. The algorithm for finding network similarities based on retweets and mentions

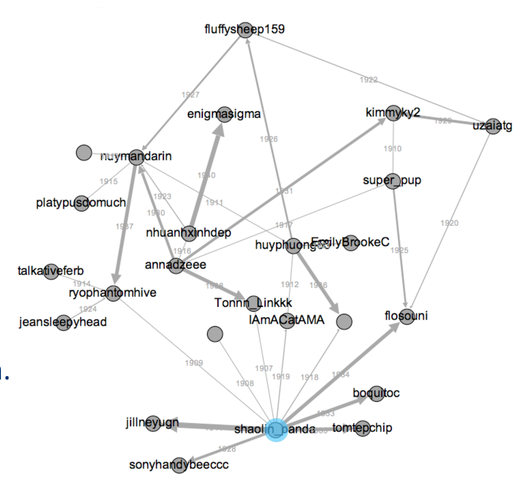

Figure 5. This is a visual representation of the network similarity algorithm concerning mentions. The edges of the graph represent number of mentions between users and have been filtered to show only the edges between 4 and 10 mentions. A user's doppelganger is most likely to be found in an area where the thickest and most number of edges are clustered, so long as the edges also point back to the user (highlighted in blue). Our analysis then takes an extra step by computing the cosine similarity within these networks to find the doppelganger.