

Identifying ‘Influencers’ on Twitter

E. Bakshy, J. Hofman, W. Mason, and D. Watt

Proceedings of the fourth ACM International Conference on Web Search and Data Mining, 2011

Bhavika Tekwani

Introduction

- Twitter has 313M monthly active users
- 1B unique page visits to sites through embedded tweets
- 31% of tweets contain a link¹
- 55% of tweets contain a photo¹
- Twitter is used for social networking, advertising, news dissipation, community building, social organizing.

References:

[1] [HubSpot - How the World Uses Twitter](#)



'Influence' in social networks

- Influence: information reaching large populations
- Potential to influence public opinion, adoption of innovations, new product market share, brand awareness
- Influencers: individuals who disproportionately affect the likelihood of information spreading broadly
- Historically, the assumption under every measure of social influence has been uniform mixing i.e., individuals interact directly.



'Influence' in social networks

- Diffusion models are now studied on networks that restrict the set of possible interactions.

- What are interactions?

Facebook: comments, likes, shares, etc

Twitter: retweets, mentions, favourites (now likes)

- Twitter has a broadcast based structure - you choose who you listen to
- Twitter provides an environment to study diffusion processes



Measuring Influence on Twitter

- Influence (user): a user's ability to post URLs which diffuse through the follower graph
- Restricted to users who “seed” content - original posters of content that they have not received through their follower graph
- Influence (tweet): no. of users who retweet the URL and this can be traced back to the original poster
- Build a model to predict influence based on the user's attributes (following, followers, activity) and evaluate the utility of this model.



Related Work

- Kwak et al. examine influence and diffusion through a comparison of 3 measures - no. of followers, page rank and no. of retweets. Ranking of influential users changes based on the measure.
- Cha et al. - no of followers, no of mentions and no of retweets. A high number of followers is not necessarily a big factor towards high influence.
- Weng et al. - compared no of followers and page rank with a modified page-rank measure accounting for topic/content. Influence depends on the measure.



Our Problem

- Predict influence, not just identify it
- Study the effect of content on diffusion
- Consider average users, not just the most influential as defined by network metrics (followers, RTs, mentions, etc)



Data

- 1.03B tweets from September 13, 2009 - November 15, 2009
- Excluded October 14-16, 2009 because of intermittent Twitter API outages
- Extracted 90M tweets which contained bit.ly URLs
- Further reduced to 87M tweets which were 'seed' URLs - posted by a user who was not following any other user who had shared the same URL.
- Subset of 74M seeds - posted by active users in both first and second months of the data collection period
- 1.6M seed users identified, each seeded 46.33 bit.ly URLs on an average



Data

- Now, crawl the follower graph for 1.6M seed users - find users who broadcast at least one URL in the same 2 month period
- Crawling this one-way directed chain of followers led to 56M users and 1.7B edges
- No of followers and following ('friends') is highly skewed
- Only active users selected - not representative of the general Twitter follower graph
- Active users tend to have more followers than the average user.



Experiment 1: Computing Influence

- Calculate the influence score for a given tweet containing a URL by tracking the diffusion of the URL from its origin through the follower graph
- This is a diffusion event or cascade. We track it till it ends.
- Temporality is a factor: if A posts the URL before B and was the only friend of B's to post it, we say A influenced B to share the URL.
- If B has more than one friend who shared the URL, we divide the credit.

Policies: primacy, split influence, last influence



Experiment 1: Computing Influence

- Create disjoint influence trees for every initial posting
- The number of users in the influence tree or cascade defines the influence score for every seed.
- Calculate the influence score based on all 3 previous measures - primacy, last influence, split influence. The values differ for all 3 measures but qualitatively, the findings are identical.
- Important: retweets and reposts are both counted as actions based on influence. This may attribute independent events to influence.



Experiment 1: Computing Influence

Limitations

- No measure of clickthrough rate on the content diffused
- Reposts are a narrow measure of influence
- Does not capture any other influence - change in purchasing behavior, political opinion, etc
- “Influencer” is defined in the context of marketing and advertising



Experiment 2: Predicting Individual Influence

How can a marketer identify an influencer to have them seed a word-of-mouth campaign?

By predicting influence!

We need to identify attributes that consistently predict influence.



Experiment 2: Predicting Individual Influence

Observations

- The distribution of cascades follows power law.
- The average cascade size is 1.14 and median is 1.
- The deepest cascades propagates 9 generations from the origin.
- Most URLs are not reposted at all.
- Even moderately sized cascades are extremely rare.



Experiment 2: Predicting Individual Influence

- Aggregate all URL posts by user , individual influence is the average size of all cascades for which the user was a seed
- Use a regression tree model with greedy optimization to predict influence score between 0 and 1.
- Folded cross-validation is used to terminate partitioning and prevent overfitting



Experiment 2: Predicting Individual Influence

- Attributes for the regression tree model were seed user attributes and their past influence
- Seed user attributes: followers, friends, tweets, date of joining
- Past influence of seed users:
 - 1) average, minimum and maximum total influence
 - 2) average, minimum and maximum local influence



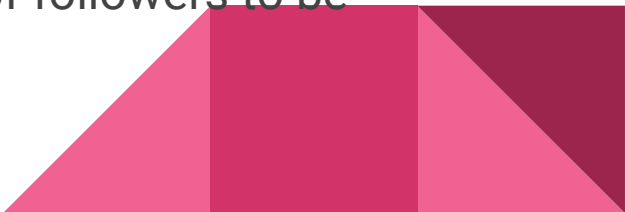
Experiment 2: Predicting Individual Influence

- Local influence: average no. of reposts by the seed user's immediate followers in the first month of observation data (train).
- Total influence: average no. of reposts over an entire cascade in the same month.
- Followers, friends, no of tweets and influence were log-transformed to account for their skewed distributions.



Experiment 2: Predicting Individual Influence

Results

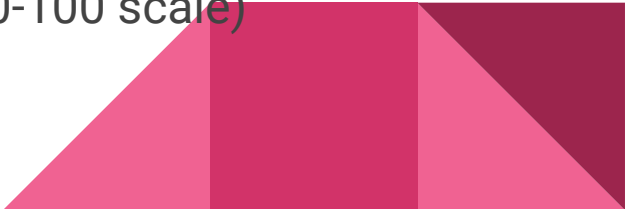
- Past local influence is the most informative measure of influence.
 - No of followers is also a good indicator of influence.
 - The prediction of the average value of influence at each division of the regression tree is quite accurate, $R^2 = 0.98$
 - However, actual model fit is poor, $R^2 = 0.34$. Large cascades are too rare for even an influencer with high local influence and no of followers to be successful.
- 

Experiment 3: The Role of Content

- Individual influence does not account for nature of content being shared.
- Assumption: certain types of content is more likely to be shared than other content - for e.g., YouTube videos > news articles
- Can we predict the cascade size based on the content & user attributes?



Experiment 3: The Role of Content

- Select 1000 URLs for this process and filter spam or non-English URLs.
 - Bin all the URLs into 10 bins, the top bin (10) contains the 100 largest cascades
 - Randomly sample 100 URLs from each of the other remaining bins.
 - Hire human classifiers from Amazon's Mechanical Turk to visit URLs and classify them as - "Spam/Not Spam/Unsure", "Media Sharing/ Social Networking, Blog/Forum, News/Mass Media or Other".
 - Ask them to rate how broadly relevant the site was (0-100 scale)
- 

Experiment 3: The Role of Content

- Rate the site based on the 'interestingness' on a 7-point Likert scale.
- Rate how positive the site made them feel on the same scale.
- Indicate whether they would share the URL on email, Twitter, IM, Facebook or Digg.
- To ensure reliability, have each URL be rated on these aspects **thrice**.
- After excluding spam or non-English links, there were 795 URLs left.



Experiment 3: The Role of Content


Observations

- Content that is more interesting tends to generate larger cascades on average and elicits positive feelings.
- Certain types of content spreads more widely, for e.g., lifestyle articles
- URLs associated with shareable media and social networking tend to spread more than URLs from news sites.



Experiment 3: The Role of Content

So are content features more useful in predicting individual influence?

- We fit a second regression tree model involving user features from the previous experiment and content features.
 - Content features are:
 1. rated interestingness
 2. perceived interesting to an average person
 3. rated positive feeling
 4. willingness to share via email, Twitter, IM, etc
 5. Indicator variables for type of URL (media/social network, blog, news/mass media)
 6. Indicator variable for content category (lifestyle, tech, sports, etc)
- 

Experiment 3: The Role of Content

Results

- On including content, the fit of the model did not improve.
- Due to the size of this train set compared to the previous one and because we're predicting the influence at the post level, there is a slight decrease in fit and calibration.
- The skewed distribution of successful diffusion events compared to the large number of failed events makes it hard to identify true successes.



Targeting Strategies

- Past local influence and no. of followers can be used to predict average future influence.
- A marketer might consider hiring several average users for a small fee as opposed to a single influencer for a large fee. Is this plausible?



Targeting Strategies

- Consider a cost function

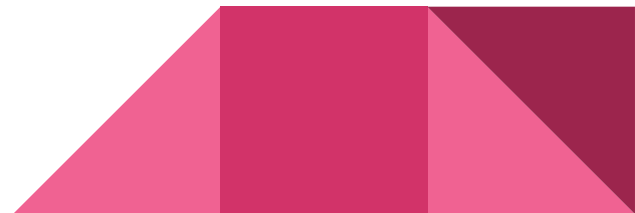
$$c_i = c_a + f_i c_f$$

c_a represents a fixed acquisition cost per individual i

c_f represents a cost per follower.

We assume $c_f = \$0.01$ based on news about paid tweets²

[2] [NYTimes: “A Friend’s Tweet Could Be an Ad”](#)



Targeting Strategies

For convenience, we write the cost function as:

$$c_a = \alpha c_f$$

When α is small, the cost function is biased towards individuals with a small number of followers - average users who are numerous

A large value of α tends to favour influencers with a large number of followers.

Since c_a varies based on α , there is a tradeoff to be made between the average users and influencers hired for seeding.

