INFO 290 - Midterm Project Report

Project title
Predicting retweet reach of a tweeter

Team Members and Roles
Natarajan Chakrapani – Topic model analysis, research, programming
Kuldeep Kapade – Data collection & analysis, programming
Karthik Reddy – Twitter network analysis, research, programming

Twitter project mentor
Stanislas Nikolov, Engineer in Search and Relevance team at Twitter, advising on Information Diffusion mechanisms

Project goals
The goal of this project is to assess the popularity of a new tweet, based on the influence source of the tweet i.e., popularity of the source tweeter, the length of the cascade chains that the tweet generated and also analysis of the content of the tweet (topics from the words or links it contains). The aim is to develop a multiple linear regression model to predict the popularity (in terms of retweet count) of a tweet, factoring in various attributes of tweets from a training set.

Project strategy
We aim to use a corpus of tweets using topic specific hashtags to, in order to get a strong topic model using LDA (Latent Dirichlet Allocation). Then we will scrape recent 200 tweets and retweets from celebrities in certain categories, and apply the topic model to these tweets to find their topic distribution. We will additionally use the follower/following counts, list count of the authors of these tweets and use it in a multiple linear regression model to predict retweet counts. Then we can use a test tweet and find which celebrity should tweet it to get max retweet.

Project timeline
November 13, 2012: Completion of first pass on the regression model.
November 25, 2012: Refined regression model.
December 4, 2012: Complete report with results visualization

Academia research for the project.

1) Meenakshi Nagarajan, Hemant Purohit, Amit Sheth. 2010. **A Qualitative Examination of Topical Tweet and Retweet Practices**. International Conference on Weblogs and Social Media (ICWSM).

http://knoesis.wright.edu/library/download/icwsm2010-cameraready-final.pdf

The appearance of sparse and dense retweet networks is dictated by the familiarity between members of that community and the narrow focus of the event which leads to high levels of user engagement. And depending on this, most content that was being retweeted, re-posted or copied, author attribution may or may not be present.

Key Takeaway - Based on this paper, we concluded that with lack of user attribution in tweets cascaded from celebrities, it would be hard to form definitive retweet chains from the source to the last tweet. It would be prudent to predict the final retweet counts of celebrity sourced tweets, rather than forming inaccurate retweet chains. We are also hampered by the limited results from the search API which does not give any results beyond the past 9 days thwarting any attempts at forming retweet chains spanning 9 days in the past.

2) Meeyoung Chay ,Hamed Haddadi, Fabricio Benevenuto, Krishna P. Gummadi. 2010. **Measuring User Influence in Twitter: The Million Follower Fallacy.** Association for the Advancement of Artificial Intelligence.
http://snap.stanford.edu/class/cs224w-readings/cha10influence.pdf

This paper debunks the myth that no. of followers is proportional to user influence. They assess three parameters associated with determining user influence - in degree, retweet and mentions. They conclude that the most connected users are not necessarily the most influential when it comes to engaging one's audience in conversations and having one's messages spread. Also, they measured user influence across topics and noted that moderate influentials like opinion leaders and evangelists also had consistent influence ranks over diverse topics.

Key Takeaway - This paper provided useful insights for our initial project idea of focussing in on celebrities/opinion leaders as our subset of users to measure relative influence.

3) Wojciech Galuba, Karl Aberer. 2010. **Outtweeting the Twitterers - Predicting Information Cascades in Microblogs.** Workshop on Online Social Network.
http://static.usenix.org/event/wosn10/tech/full_papers/Galuba.pdf

This paper describes the propagation of the URLs in the Twitter network based on the content popularity, user influence and the rate of propagation.

Key Takeaway - We plan to use a similar idea and extend the information cascade analysis based on both content (topic model using LDA) and the user influence of the tweeter.

4) Roja Bandari, Sitaram Asury and Bernardo Hubermanz. **The Pulse of News in Social Media: Forecasting Popularity.**
http://www.hpl.hp.com/research/scl/papers/newsprediction/pulse.pdf

In the paper, the objective was to design features based on content to predict the number of tweets for a given news article. Topics for these articles were pre-tagged from news sources.

Key Takeaway - This paper focused on textual attributes to determine the popularity of the news items. We aim to find popularity of tweet content and combine this with the tweeter characteristics regarding influence that can affect the virality of the tweet cascade.

5) Bakshy E, Hofman M J, Wason A. M, Watts J. D. 2011. **Everyone's an Influencer: Quantifying Influence on Twitter.** WSDM
http://misc.si.umich.edu/media/papers/wsdm333w-bakshy.pdf
This paper supports the idea that influence of a tweeter is determined by the number of his/her followers reposting the same tweet rather than just counting the number of followers only. The cascade size increases with time and this rate also has to be taken into account when predicting the popularity.
Key Takeaway - This paper provided validation to our project idea that , content of a tweet along with source tweeter credentials/influence is important in increasing the reach of a tweet in terms of retweet count. Topic models of tweets generated by celebrities might hold clues to what a potential marketer could target and incentivize the celebrities to tweet for maximal diffusion.

6) Jake Lussier and Jacob Bank. 2011.**Final Report: Local Structure and Evolution for Cascade Prediction.**
http://snap.stanford.edu/class/cs224w-2011/proj/jbank_Finalwriteup_v1.pdf
This paper presents an idea to predict cascade size growth which could be useful when calculating the influence of tweeter. They primarily use a supervised classification scheme to predict if cascade size will exceed 20, given a size 10 or less sized cascade network. They showed that most retweets follow shortly behind the original tweet, and found little relation between degrees of nodes in a cascade and the size of that cascade
Takeaway for project : This sparsity of retweet chains, helped us focus on final retweet counts as a reliable measure of reach (lower bound). We aim to predict the final retweet counts , given the topic distribution of tweets and the follower/following ratio of the source tweeter.

7) Liangjie Hong Ovidiu Dan Brian D. Davison. 2011. **Predicting Popular Messages in Twitter.** World Wide Web (WWW)
http://www.cse.lehigh.edu/~brian/pubs/2011/WWW/predicting-popular-messages-twitter.pdf
This paper takes the problem of predicting the popularity of messages into
1) A binary classification problem that predicts whether or not a message will be retweeted, and,
2) A multi-class classification problem that predicts the volume of retweets a particular message will receive in the near future.
Here, the number of messages following a specific message is treated as the number of retweets this message will receive. Apart from the obvious conclusion of effect on user in degree on retweet probabilities, temporal features have a stronger effect on these messages with low and medium volume of retweets, compared to highly popular messages
Key Takeaway - We aim to extend this and utilize the 'retweet count' of tweets to see the final reach of a tweet, given a user.

8) Sasa Petrovic , Miles Osborne , Victor Lavrenko. 2011. **RT to Win! Predicting Message Propagation in Twitter.** International Conference on Weblogs and Social Media (ICWSM).

http://homepages.inf.ed.ac.uk/miles/papers/icwsm11.pdf
The paper is focussed on prediction of retweets on streaming tweets, and hence is more real-time in nature. In order to have some accuracy, they utilise the social features and tweet features to build a binary classifier. Notable results mention that user features like #followers,#friends and listed are important features for this classification and #mentions, #statuses impact the classification task adversely.
Key Takeaway - we aim to utilize the follower/following ratio of the celebrities to incorporate the retweet count prediction.

9) Xufei Wang, Huan Liu, Peng Zhang, Baoxin Li. **Identifying Information Spreaders in Twitter Follower Networks**
http://dmml.asu.edu/users/xufei/Papers/TR-12-001.pdf
Rather than retweet prediction or understanding information diffusion , this paper tries to answer who spreads or relays new ideas in social network. Backed up by huge datasets, They had an insightful take on the reasons why retweet history might not be a good bet retweet predictions - since users rarely retweet consistently and no. of your followers who retweet you consistently are also insignificant . This is true for a majority of twitter users (celebrities are an exception).
Key Takeaway - This paper helped solidify our idea that we should focus on opinion leaders / mini-celebrities to understand the measure of their influence(in terms of retweet counts).

Accomplishments to date on the project

Further refinement of project goals
Since, the inception of the project our project goal has been refined a lot after talking to various people and with the help of our advisor. We are simplifying the process of finding the influencers from the corpus of users. We will be able to identify the right target influencing user for the marketer which will provide the best probability for a particular tweet to go popular. We use LDA based topic modelling to identify topic classification for particular tweet and then use to match relevance of the tweets by influencers. Then we also use linear regression to predict the tweet popularity of the tweet.

The basic idea is a marketer will login to our product and just provide the tweet text they want to promote, and then our system will produce of the ranking of suitable users who are best suited for the marketer to pursue to promote their tweet.

Data gathering
As of now we have gathered tweet, retweet and followers count data from handpicked diverse set of users. These users are supposed influencers in various topics. We have also added few average users to verify our results at the end. To train our LDA, we fed data from topic based hashtags from around 10-12 topics and also few known random tweets from familiar topics. Our basic idea at this stage is to train the algorithm using predictable data so we can judge the accuracy of

the algorithm even by analysing it manually. Once, we gain the confidence on the algorithm we will be able to run it easily on any random data and expect accuracy.

Algorithm development

Use LDA (Latent Dirichlet allocation) for topic modelling on tweets gathered by topical hashtags (e.g #technology, #business etc). We identify 10 topics over the distribution of words in the corpus and determine the distribution of topics for new tweets. Each new tweet is assigned a score for each topic, with the sum of all scores being one. The higher the score for a topic, the stronger that tweet reflects that topic content. Each celebrity tweet will comprise a vectors of size 10 having scores for each topic. Additionally each tweet will also use the follower/following user ratio and the listed count attribute to link with the creator of the tweet (celebrity). These 12 variables will be the explanatory variables to be used in a multiple linear regression model with the retweet count associated with each tweet being the response variable.

Multiple linear regression is appropriate here because we expect the topics provide by LDA and the other features like friend/following ratio and listed count to be not strongly correlated, thereby adding improvement to the prediction accuracy of the retweet count.

Software architecting

Our project design is divided into three parts - data gathering module, data training module and marketer interface which will take tweet as data input. Theoretically first two modules won't run frequently, they can be updated only on weekly basis. They can be written as a cron job on a system. The interface module will act as front-end to the marketer (ideally a web application) and can be run on regular basis to help marketer in predicting popularity of the tweet and identify right promoter for the given tweet.

Coding

Currently data gathering module is written in Java, which is complete at this point. Data training module is written Python which uses LDA and linear regression on gathered training data. This module is almost nearing completion. We started a bit on front-end module, but this will mainly come into picture once we gain confidence with our algorithm's accuracy.

Interface design

Our front-end will be a web application where marketer will login to gain analysis on their prospective tweets. We also plan to use tree based graphs help better visualize rankings of promoters and popularity prediction of the tweets.

Next steps

Our plan is keep iterating on our algorithm until we find good accuracy. We can experiment with various models and try to answer other questions for the marketer such as predict the retweet probability for a given user. Our focus will also be on creating good interface for the marketer to help them visualize and analyze various predictions we plan to provide.

Work percentages by each team member

Task	Natarajan	Karthik	Kuldeep
Research	35%	35%	30%
Data Gathering	25%	35%	40%
Algorithm Design	33%	33%	33%
Documentation	45%	25%	35%