PARATWEET

A TWITTER CONTENT BASED RECOMMENDATION ENGINE

ROHIT TURUMELLA, ANTHONY SALGADO, SANKETH KATTA, JAMIE TURLEY

Table of Contents

Team Members and Mentor

1. Rohit Turumella
2. Anthony Salgado
3. Sanketh Katta
4. Jamie Turley
5. Shai Haim – Mentor

Project Background and Motivation

The goal of ParaTweet is to create a new recommendation engine for Twitter users that gives them suggestions for who to follow based on the content that one consumes.

There are a large number of users on Twitter and a common complaint amongst many users of the service is that it is hard to find new users to follow who are similar to the people that they follow. Although Twitter has a few features that deliver suggestions for users to follow and has curated lists for certain topics, the current approaches are centered on closing the links in one's social graph (Triadic Closure) and creating static lists for various topics. In our conversations with our mentor, we came to realize that the current way that Twitter recommends people to follow ignores an important facet of Twitter – one's user timeline (which contains all the tweets of the people that they follow).

Follow and List Recommendation Engines are extremely valuable to the Twitter ecosystem because they help generate growth, increase user engagement with the service, and reduces churn by keeping users engaged with new content that is relevant to them.

ParaTweet generates recommendations based on the content that a user consumes which allows us to conduct textual analysis to get a more holistic understanding of the type of content that a user prefers. The recommendation engine functions as a web application, which is currently released to the public (http://twitter190.cloudapp.net) that allows users to enter a given Twitter Username and get a list of personalized recommendations for users that that person should follow.

Project History and Pivot

The team initially started out with trying to design an application which would analyze a Twitter user's public tweets and give them a professionalism score by determining how similar their tweets were to the tweets of chosen industry leaders. These industry leaders would be chosen by pulling the users with the highest Klout scores in a specific category. Users would have been able to choose from a list of industries and receive a "Professionalism Score" for that industry. The motivation for this was because the team thought it would be useful for job seekers who need to maintain a modicum of professionalism on social networking platforms and to allow recruiters to quickly gauge the professionalism of job applicants.

After receiving some valuable feedback from our mentor and our professor and a detailed literature review, we realized that professionalism was something that was hard to quantify because it is a very subjective topic. We took a look at the codebase and after talking with our mentor, who commended

us on our technical approach and the fact that we had a working tweet similarity analyzer with excellent results, our mentor provided us with a suggestion for how we could leverage our existing technical infrastructure and classification algorithms to achieve something that would be more technically interesting and also create a lot of value for the Twitter ecosystem. Shai suggested that we take our classification algorithms to conduct textual analysis on the tweets that a user consumes and provide them with suggestions for people to follow which is how ParaTweet came to be.

Revised Project Timeline and Goals

Despite the pivot in our project, the team accomplished many of the goals we committed to. The algorithm we designed accurately computes the similarity between Twitter users' public timelines. This is evinced by the fact that one can run our application on one of our "ground truth users" from which we take our recommendations and clearly see that the user has been correctly clustered with users who produce similar content. For example, if one were to run the algorithm on username "barackobama", the results of the similarity algorithm include users like "whitehouse", "abcnews", "wsj" who often produce content relating to the field of politics. This functionality is similar to the "Similar to … " feature on Twitter.

In addition to this functionality, one can use our application to generate a list of recommendations from our corpus of top users provided that the username entered is not a top user. Our algorithm's accuracy is illustrated by the fact that the user is already following many of the users generated by our algorithm. Ultimately, this proves the validity of our content-similarity-based recommendation algorithm.

The algorithm required little fine-tuning, but reverted back to using pure term-frequency instead of TF-IDF weighting for our cosine-similarity implementation.

The group exceeded the goal of having an interesting and intuitive user-experience. Twitter Bootstrap helped make the front-end aesthetically pleasing. Also, using a combination of AJAX requests and aggressive caching helped make a user's actual wait-time between request and service much lower than expected.

To use our application, a user sees an input box where s/he may a username. If the result is pre-cached, then the textbox will auto-populate to facilitate frequent visitors. Otherwise, a user may enter a valid Twitter username and wait for their results. The list will appear via AJAX right below the input box. For future visits, the user's recommendations will be cached for instantaneous results.

Technical Stack and Code

The application was developed on the following technical stack, which was modeled after Instagram's Technical Stack:

1. On the frontend, we primarily used Twitter Bootstrap (http://twitter.github.com/bootstrap/), a front-end development framework, and JQuery (http://jquery.com/)

2. On the backend, the application was built in Python and Flask (http://flask.pocoo.org/) was used as the web development framework because it is a more light weight web development framework and allowed us to develop the application faster. It also featured the Jinja HTML-template library (http://jinja.pocoo.org/docs/)
    a. For our database, we chose MongoDB (http://www.mongodb.org/) as our DBMS of choice because it allows for a more flexible schema compared to relational DBMS's, we were doing a large amount of batch processing with regards to scraping and aggregating tweets, and large part of the algorithm depended on matching documents.
3. For the aggregation and collection of tweets, the team chose to use Python-Twitter (http://code.google.com/p/python-twitter/) to interact with the Twitter API, the Python Stemming library (http://pypi.python.org/pypi/stemming/1.0) to conduct Porter Stemming, and the NLTK (http://www.nltk.org/) library to remove stopwords and conduct textual analysis.
4. Because the stack was not a standard stack supported by Heroku, we provisioned a Virtual Machine on Windows Azure (a public cloud service) and set it up to scale to large amounts of users.
    a. Green Unicorn (http://gunicorn.org/) was used for the WSGI server with multiple workers to allow the app to support multiple concurrent long running calculations while still being able to handle additional requests.
    b. Nginx (http://wiki.nginx.org/Main) was used for the HTTP server in front of Green Unicorn to handle static file serving, without having static requests reach the application servers, and as a reverse proxy to forward appropriate requests to the Flask app via Green Unicorn.
    c. Supervisor (http://supervisord.org/) was used for Daemon management, to control (stop, start, reload) the multiple Green Unicorn processes easily.

Code: The code is currently hosted on GitHub on a private repository and can be accessed at https://github.com/sankethkatta/tpro.

Similarity Algorithm

In order to generate the suggestions for people to follow, we compared the tweets of a user's followers to the corpus of tweets that we gathered from the top 1000 Twitter users with the most followers through the cosine similarity algorithm.

Other approaches included TF-IDF and the Jaccard Coefficient but with our application we noticed better performance and recommendations with Cosine Similarity.

However, because of the given vector size, any similarity algorithm takes a large amount of time to compute. As a result, the recommendation engine is not meant not to run in real time (takes 3-5 minutes to generate a report for a fresh user and the application caches recommendations for previously computed users for a few weeks). This is a reasonable expectation because recommendation engines at any company (LinkedIn's People You May Know recommendation engine, Twitter's Who to Follow recommendation engine, Netflix's Movies You May like

recommendation engine, etc.) are not meant to be run in real time; in reality, they are run as a background process and each user has pre-generated recommendations which are periodically refreshed.

Future Work

The team was pleased with the results of the project and the positive feedback that we received from Twitter Engineers and users of the application. In the future, the team plans to:

1. Parallelize and optimize the scraping and cosine similarity algorithm.
2. Create a script which conducts a Breadth-First Traversal of all Twitter users to ensure that users do not have to wait for a few minutes to receive a report for the first time.
3. Introduce support for foreign languages:
    a. Enhance the current corpus of data with notable Twitter users from different countries.
    b. Utilize IP Geolocation to facilitate location-based recommendations and weight users who are from their country over other users in our recommendations.

Working Application

The web application is currently in production and can be accessed at http://twitter190.cloudapp.net (using the application is self-explanatory).

Work Allocation

	Rohit Turumella	Anthony Salgado	Jamie Turley	Sanketh Katta
Back End – Scraper, Corpus Aggregation, Misc	70	15	15	
Back End- Similarity		100		
Front End			30	70
Literature Review	33		33	33
Meetings	25	25	25	25
Midterm Report + Presentation	20	20	40	20
Final Report + Presentation	60	20	10	10
Scaling + Architecture + DevOps + Misc	20	20	20	40
Application Testing + Fixing Performance Issues	20	40	20	20

Literature Review

1. **Anger, Isabel, and Christian Kittl. "Measuring influence on Twitter." Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies. ACM, 2011.**
    a. The article explores the methods in measuring social influence among individual Twitter users. Anger and Kittl compare these methodologies using the top 10 Twitter users in Austria in order to show the drawbacks and benefits of each.
    b. Klout - Klout measures, as it states on its website, a user's overall online influence with a score ranging from 1 to 100, with 100 being the highest amount of possible influence. Klout analyses more than 25 variables, also offering the possibility to combine the scores from all three analyzed platforms.
    c. Twitter Grader - which calculates a score out of 100. Also kept secret, however it is communicated that considered factors include number of followers, Twitter Grader score of followers, number of tweets, update recent-cy, follower/following ratio, and engagement, (i.e retweet and mention ratio).
    d. Ultimate conclusion: every approach is different from the others in terms of algorithm and emphases on different individual factors, thereby resulting in different rankings of

sample users. This is due to the fact that there is no consent on what indicates influence on Twitter.

e. Researchers developed something called SNP - Social Networking Potential
    i. Number of followers, of individual interactors, retweets, mentions, and total amount of tweets.

2. **Weng, Jianshu, et al. "Twitterrank: finding Topic-Sensitive Influential Twitterers."Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010.**
    a. Researchers propose a new way measure social influence in Twitter, called TwitterRank.
    b. Motivation: "reciprocity" is a very prevalent phenomenon in Twitter.
        i. 72.4% of users in Twitter follow more than 80% of their followers
        ii. 80.5% of the users have 80% of users they are following follow them back
    c. However, researchers point to something called homophily, which show that there are a number of serious Twitter followers on the web. (e.g. Not all users randomly "follow" on the web)
    d. TwitterRank - an extension of Google's PageRank algorithm that measures the influence taking both the topical similarity between users and the link structure into account.
3. **Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010, May). Measuring user influence in twitter: The million follower fallacy. In 4th international aaai conference on weblogs and social media (icwsm) (Vol. 14, No. 1, p. 8).**
    a. Compares three measures of influence:
        i. Indegree influence - the number of followers of a user, directly indicates the size of the audience for that user.
        ii. Retweets - indicates the ability of a particular user to generate content with pass-along value.
        iii. Mentions - indicates the ability of that user to engage others in a conversation.
    b. Researchers have found that popular users who have high Indegree are not necessarily influential in the context of spurring retweets or mentions.
    c. Most influential users can hold significant influence over a variety of topics
    d. Influence is not gained instantly or accidentally, but through concerted efforts such as limiting tweets to a single topic.
    e. Contradicts (Wang et al.) by observing that their more complete data set has low reciprocity. And instead predicts that social links on Twitter represent an influence relationship, rather than homophily.
    f. Examined the users who increased their influence over a short period of time to answer the question: "what behaviors make ordinary users influential?"
        i. Manual inspection revealed that users who limit their tweets to a single topic showed the largest increase in their influential scores

4. **Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008, February). Finding high-quality content in social media. In Proceedings of the international conference on Web search and web data mining (pp. 183-194). ACM.**
    a. Researchers investigate methods for exploiting community feedback to automatically identify high quality content.
    b. Research Subject: Yahoo Answers, a large portal that is particularly rich in the amount and types of content and social interaction available in it.
        i. What are the elements of social media that can be used to facilitate automated discovery of high-quality content?
        ii. How are these different factors related? Is content alone enough for identifying high-quality items?
        iii. Can community feedback approximate judgments of specialists?
    c. What was interesting to us: researchers measured punctuation, grammar, as well as typos, which can be considered "professional". Which is something that we decided not to do.
    d. Presented a general classification framework for quality estimates in social media: graph-based model of contributor relationships combined with content/usage based features.
5. **Bakshy, Eytan, et al. "Everyone's an influencer: quantifying influence on twitter." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.**
    a. In this paper researchers investigate the attributes and relative influence of 1.6M Twitter users by tracking 74 million interactions.
    b. Conclude unsurprisingly that the largest "cascades" or interaction ripple effects show in the media-sphere are generated by users who have been influential in the past and who have a large number of followers.
    c. Crawled the portion of follower graph compromising all users who had broadcast at least one bit.ly URL over the two-month period.
    d. What was interesting: URLs that were rated more interesting and/or elicited more positive feelings by workers were more likely to be passed along if user frequently tweeted about a particular topic.
6. **Cataldi, Mario, Luigi Di Caro, and Claudio Schifanella. "Emerging topic detection on Twitter based on temporal and social terms evaluation." Proceedings of the Tenth International Workshop on Multimedia Data Mining. ACM, 2010.**
    a. Researchers propose a topic detection technique that outputs real-time relevant topics. As our definition of professionalism is dependent on the industry, this might be an interesting technique for us to look into.

b. Algorithm first extracts the set of terms of the tweets and model the term life cycle according to something the researchers define as “emerging theory”.
    i. A term can be defined as emerging if it frequently occurs in the specified time interval and it was relatively rare in the past.
c. Determines authority of users using the Page Rank algorithm.
d. Concept of Content Nutrition
    i. Different tweets containing the same keyword generate different amount of “nutrition”--e.g. quality of tweet--depending on the representativeness of the author in the considered community.
e. Uses a word co-occurring algorithm that includes a correlation vector that contains the relevant keyword for a potential topic.

7. **Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on twitter. In Proceedings of the 20th international conference on World Wide Web (pp. 675-684). ACM.**
    a. The article explores the issue of information credibility on Twitter which has become a very important issue lately because Twitter is now the fastest mechanism through which news disseminates. Although most messages are truthful, many people abuse the service by spreading false rumors and taking advantage of the network effect to spread false news.
    b. The researchers have tried to figure out a mechanism by which they can automatically assess the credibility of a tweet by looking at other trending tweets with the same content or hashtag, the poster’s retweeting behavior, textual analysis, and whether it links to some other content.
        i. This is relevant to us because part of our definition of professionalism is whether a given user is credible or not. This is because in the future we envision our tool being used by recruiters and other users of the service to determine whether the user is credible to follow or hire.
    c. In a user study, it was determined that in the absence of a quantitative measure of credibility, people determine credibility in very subjective ways such as gender, design of the profile and profile picture.
    d. The researchers used Twitter Monitor over a 2 month period to find random bursts of activity, used Mechanical Turk to classify a selected set of tweets as news or chat, and built a classifier that used ground truths from mechanical turk users rating a certain set of tweets as credible or not credible.
        i. It took advantage of examining the replies to the tweet, profile metrics of the tweet author (number of tweets, how old the account was, number of followers, retweets), and looking at URLs in tweets

1. This could be a possible area of future expansion for our classifier which currently uses cosine similarity

e. Results show that it is possible to separate credible content in a rapid manner

8. **Johnson, K. A. (2011). The effect of Twitter posts on students' perceptions of instructor credibility. Learning, Media and Technology, 36(1), 21-38.**
    a. Article aims to examine the effect of Twitter posts on the perceived credibility of teachers by their students
    b. Aimed to examine whether credibility was affected by social content or links to professional content
        i. Broke up teachers into three groups
            1. Teachers who tweet solely professional content
            2. Teachers who tweet solely social content
            3. Teachers who tweet a mix of social and professional content
        ii. Had Students rate these teacher accounts based on their perceived credibility
    c. Results were surprising: Students rated the teachers who tweeted only social content higher than the teachers who only tweeted professional content
        i. Important for us because we initially thought of defining our metric solely on professionalism and made us examine whether we should add a social metric component to it
    d. Most users never click on hyperlinks in tweets
9. **Burstein, J., & Wolska, M. (2003, April). Toward evaluation of writing style: finding overly repetitive word use in student essays. In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1 (pp. 35-42). Association for Computational Linguistics.**
    a. The essay explores the issue of automated essay scoring and particularly looks at a commercial technology called Criterion which uses ML techniques to find repetitive word use in student essays
        i. Relevant because professionals use good writing style
    b. Essay grading technologies depend on ML techniques to analyze text and model kinds of analysis
        i. As a result, depends on large corpus of essay data. Kind of like us where we are depending on a large corpus of tweets from people Klout views as influencers
    c. Had 2 people go through the corpus of essays and mark what they viewed as repetitive behavior
        i. Through this process they found 7 vectors which reliably predicts whether a student's essay was repetitive
            1. The Seven Vectors were: Absolute Count, Essay Ratio (ratio of repeated words to total essay word count), Paragraph Ratio (average occurrence of

the word in a paragraph), Highest Paragraph Ratio, Word Length, a boolean value for whether the word was a pronoun, distance between the word and its previous occurrence

d. Even though this is a very subjective measure, paper shows that it's possible to build an automated system which recognizes whether students are writing essays which don't vary in vocabulary

10. **Juffinger, A., Granitzer, M., & Lex, E. (2009, April). Blog credibility ranking by exploiting verified content. In Proceedings of the 3rd workshop on Information credibility on the web (pp. 51-58). ACM.**
    a. The article aims to create criteria to rank the credibility of blogs. Twitter, being often referred to as a "micro-blogging" service is relevant to the analysis.
    b. They have 2 main criteria for ranking each blog's credibility, one being a comparison of content similarity to that of a "verified news corpus".
        i. Our app uses top Klout Influencers as the "verified corpus"
    c. People tend to be less controlled on blogs as are identified by nothing more than a username. Structure of the blog itself is difficult to use to judge its credibility (unlike a traditional website) as most have similar or identical structures.
        i. Twitter faces much the same problem, each user's twitter profile is identical, making the structure of the page impossible to use as criteria.
    d. Centroid Cosine similarity is used to rank the content of each blog against the verified corpus.
11. **Ziegler, C. N., & Lausen, G. (2004). Analyzing correlation between trust and user similarity in online communities. Trust Management, 251-265.**
    a. This article looks at the correlation between user trust and similarity.
    b. The analysis follow the "All Consuming book-reading community", an online book reading community
        i. All Consuming already has a defined way to assign relations between trusted users.
        ii. Similarity was assigned based on category descriptors for book ISBNs on Amazon.
        iii. Each Amazon taxonomy could further point to super-topics (Matrix Analysis can relate to Algebra)
    c. In order to compute the correlation, they mention both Pearson's Correlation Coefficient and Cosine Similarity as popular choice.
        i. They opt for Pearson's because of its ability to compute negative correlation.
        ii. It allowed them to categorize users who are strongly diverging in topics.
    d. Users were found to be 23% more similar to their trusted connections than to any arbitrary user.

e. Leveraging the fact that users are more similar to their trusted connections, we can look at our similarity ranking in reverse. Top influencers would be significantly more similar to trusted connections than a random users. A user who has a high similarity score would put them in the same category as trusted users. We would assume that the trusted connection of a top influencer would be a "professional" one.

12. **Maia, M., Almeida, J., & Almeida, V. (2008, April). Identifying user behavior in online social networks. In Proceedings of the 1st workshop on Social network systems (pp. 1-6). ACM.**
    a. This article looks at the correlation between user trust and similarity.
    b. The analysis follow the "All Consuming book-reading community", an online book reading community
        i. All Consuming already has a defined way to assign relations between trusted users.
        ii. Similarity was assigned based on category descriptors for book ISBNs on Amazon.
        iii. Each Amazon taxonomy could further point to super-topics (Matrix Analysis can relate to Algebra)
    c. In order to compute the correlation, they mention both Pearson's Correlation Coefficient and Cosine Similarity as popular choice.
        i. They opt for Pearson's because of its ability to compute negative correlation.
        ii. It allowed them to categorize users who are strongly diverging in topics.
    d. Users were found to be 23% more similar to their trusted connections than to any arbitrary user.
    e. Leveraging the fact that users are more similar to their trusted connections, we can look at our similarity ranking in reverse. Top influencers would be significantly more similar to trusted connections than a random users. A user who has a high similarity score would put them in the same category as trusted users. We would assume that the trusted connection of a top influencer would be a "professional" one.

Bibliography

1. Anger, I., & Kittl, C. (2011, September). Measuring Influence on Twitter. In Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (p. 31). ACM.
2. Weng, Jianshu, et al. "Twitterrank: finding Topic-Sensitive Influential Twitterers."Proceedings of the third ACM international conference on Web search and data mining. ACM, 2010.
3. Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010, May). Measuring User Influence In Twitter: The million follower fallacy. In 4th international aaai conference on weblogs and social media (icwsm) (Vol. 14, No. 1, p. 8).
4. Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008, February). Finding high-quality content in social media. In Proceedings of the international conference on Web search and web data mining (pp. 183-194). ACM.
5. Bakshy, Eytan, et al. "Everyone's an influencer: quantifying influence on twitter." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.
6. Cataldi, Mario, Luigi Di Caro, and Claudio Schifanella. "Emerging topic detection on Twitter based on temporal and social terms evaluation." Proceedings of the Tenth International Workshop on Multimedia Data Mining. ACM, 2010.
7. Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on twitter. In Proceedings of the 20th international conference on World Wide Web (pp. 675-684). ACM.
8. Johnson, K. A. (2011). The effect of Twitter posts on students' perceptions of instructor credibility. Learning, Media and Technology, 36(1), 21-38.
9. Burstein, J., & Wolska, M. (2003, April). Toward evaluation of writing style: finding overly repetitive word use in student essays. In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1 (pp. 35-42). Association for Computational Linguistics.
10. Juffinger, A., Granitzer, M., & Lex, E. (2009, April). Blog credibility ranking by exploiting verified content. In Proceedings of the 3rd workshop on Information credibility on the web (pp. 51-58). ACM.
11. Ziegler, C. N., & Lausen, G. (2004). Analyzing correlation between trust and user similarity in online communities. Trust Management, 251-265.
12. Maia, M., Almeida, J., & Almeida, V. (2008, April). Identifying user behavior in online social networks. In Proceedings of the 1st workshop on Social network systems (pp. 1-6). ACM.