Twitter Professional Analysis Tool

Big Data with Twitter – Info 290

Team Members

1. Rohit Turumella
2. Anthony Salgado
3. Sanketh Katta
4. Jamie Turley

Name of Twitter Project Mentor

- Shai Haim

Project Goals

The goal of this project is to create an application that will analyze a Twitter user's public tweets and determine the similarity between those tweets and the aggregated tweets of industry leaders. Industry Leaders will be chosen by pulling the users with the highest Klout scores in a specific category that we are trying to analyze. Users will be able to choose from a list of industries and receive a similarity score. This is useful to people such as job-seekers who need to maintain a level of professionalism on their social-networking platforms.

Strategy

- *Rohit Turumella & Anthony Salgado* – backend development in addition to designing the database models and ensuring a resilient application
- *Sanketh Katta & Jamie Turley* - front-end development of the product and its connection with the backend.

Project Timeline

- *November 9th*
    a. Each team member will review documents relevant to the application and discuss their findings with the group.
- *November 13th*
    a. Database schema well defined.
    b. Code in place to retrieve relevant information from Twitter. The database will then be populated by the results of the information retrieval code.
    c. First pass at UX will be mocked up and a front-end to do simple verification of results.
    d. Algorithm for determining similarity finalized, to be implemented in the second phase.
- *December 3rd*
    a. First pass at similarity algorithm and UX implemented

    b. Prioritize and fix bugs in the application, fine-tune algorithm if unsatisfactory results.

- *December 3rd – 10th (Testing)*
    a. Documentation for the application
    b. Application thoroughly testing with no obvious bugs. UX refined and finalized after getting user-feedback.

Literature Review

For the literature review, we investigated scholarly articles that we thought were relevant to our project, as well articles that shared ideas that we could possibly entertain by the end of the semester. The following are in bibliographical order:

- **Measuring Influence on Twitter**
    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.

- **TwitterRank: Finding topic-sensitive influential twitterers**
    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.

- **Measuring Social Influence on Twitter**
    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 thay 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

- **Finding High Quality Content in Social Media**
    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.

- **Everyone's an influencer: quantifying influence on Twitter**
    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.

- **Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation**

  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-occuring algorithm that includes a correlation vector that contains the relevant keyword for a potential topic.

- **Information Credibility on Twitter**

  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

- **Toward evaluation of writing style: finding overly repetitive word use in student essays**

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 its possible to build an automated system which recognizes whether students are writing essays which don't vary in vocabulary

- **The effect of Twitter posts on students' perceptions of instructor credibility**

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

- **Blog Credibility Ranking by Exploiting Verified Content**
    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.
        i. Our app similarly uses a given user’s tweet’s as their “blog content”

- **Analyzing Correlation between Trust and User Similarity in Online Communities**
    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.

- **Identifying user behavior in online social networks**
    a. This article aims to analyze the behaviour of users on a social network. Online interactions are much more complicated, because individual categorizations are not sufficient.
    b. They recorded the usage patterns of users on YouTube. There are number of activities a user can choose to do while on the site, all contributing to the pattern.
    c. The users were then clustered together using K-means. 5 different clusters of users emerged. Each cluster signified a different type of usage pattern.
        i. Examples were *Content Producer, Content Consumer.*
    d. For our analysis, we can try and cluster users together based on similarity instead of usage patterns. We hope to see clusters of content based on the industry. With cosine similarity we would expect to see that clustering all industry leaders based on content we would naturally see industries form. (All top political influencers would cluster together

Accomplishments to Date

- Scraper code to generate csv files for the top 10 selected influencers.
- Running web application that takes in any arbitrary string and returns tables of cosine similarities of industry leaders’ tweets.
- Application will be demoed during the presentation.

Work Allocation

	Rohit Trulumella	Anthony Salgado	Sanketh Katta	Jamie Turley
Back End - Scraper	100			
Back End - Similarity		100		
Front End			100	
Literature Review	25		25	50
Meetings	25	25	25	25
Presentation & Report	20	20	20	40

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, J., Lim, E. P., Jiang, J., & He, Q. (2010, February). Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 261-270). ACM.
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.