Twitter Analyses

Laura Wiley
Wednesday, October 08, 2014

Summary

First to get us started, just a few summary statistics on the tweets:

Year	Tweeter Count	Tweet Count
2011	249	1447
2012	496	4074
2013	628	5333

Based on the latest copy of the draft comments, one of the major opinion/conclusions of twitter was:

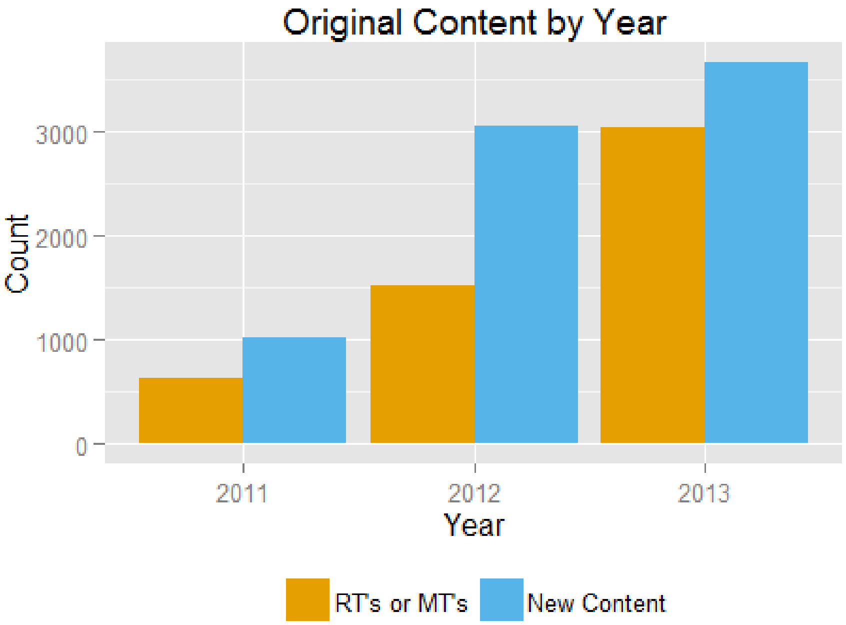
> "Our analysis of Twitter data from the AMIA 2011-2013 Fall Symposia highlighted several different usage patterns for social media in a conference setting. Usage patterns included: sharing information with those who cannot attend the conference, enhancing the conference experience for those in attendance, and building professional relationships with other conference attendees and the wider biomedical informatics community."

Given this thought I upgraded our analyses from simple hashtag output to a more meaningful syntactic analysis.

Content Types

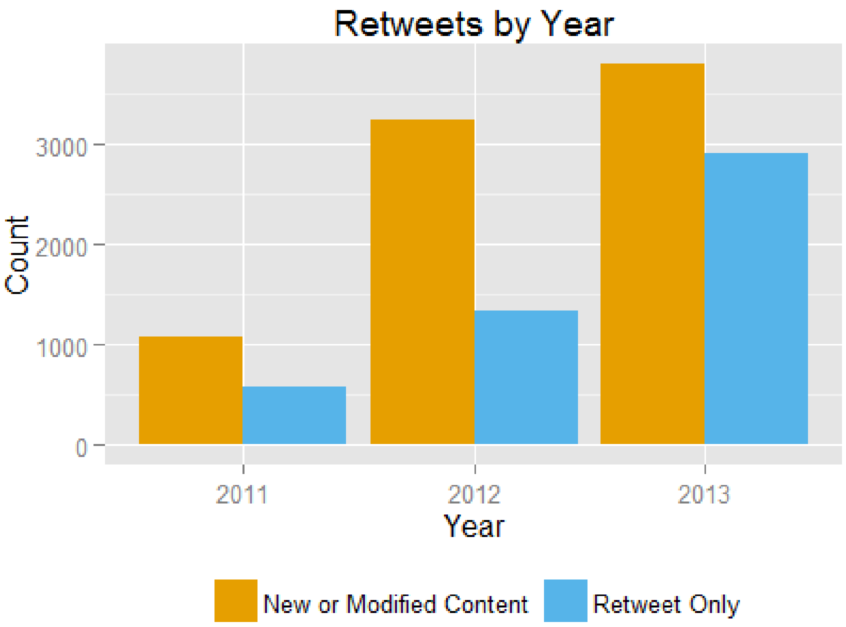
I thought it would be helpful to see how the content distribution (RT, MT, or new content) has changed as tweeting at conferences has become more common.

The first thing I was interested in was how much original content was created compared to any type or RT or MT. I have noticed an increased proportion of RTs each conference and I was curious to see if the data supported that trend. I think this also has strong implications on reach of the message in each symposium. Each RT indicates a wider network exposed to that content.



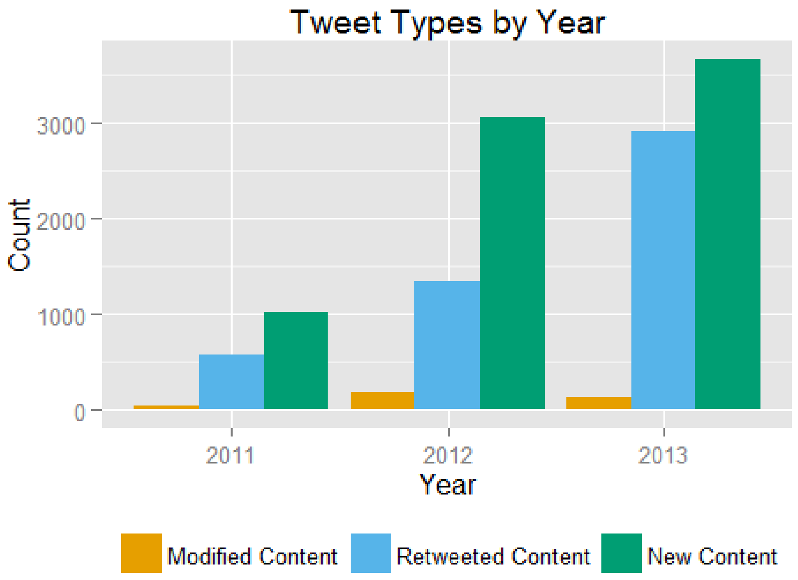
Year	RT's or MT's	New Content
2011	629	1020
2012	1528	3060
2013	3048	3671

However, I also thought that RT/MTs are really different types of tweets. A pure RT (i.e., no additional commentary) is purely sharing, whereas MTs or RTs with content before the RT are conversation/commentary. I also have noticed a large increase in RT frequency at recent meetings. I wanted to see if these impressions were supported by the data.



Year	New or Modified Content	Retweet Only
2011	1071	578
2012	3240	1348
2013	3805	2914

To pull both of these ideas together, I created three classifications of tweet type: 1. New Content 2. RT or MTs with additional thoughts added 3. Pure RT I plotted these by year as well to get a sense of how the distribution has changed over time.

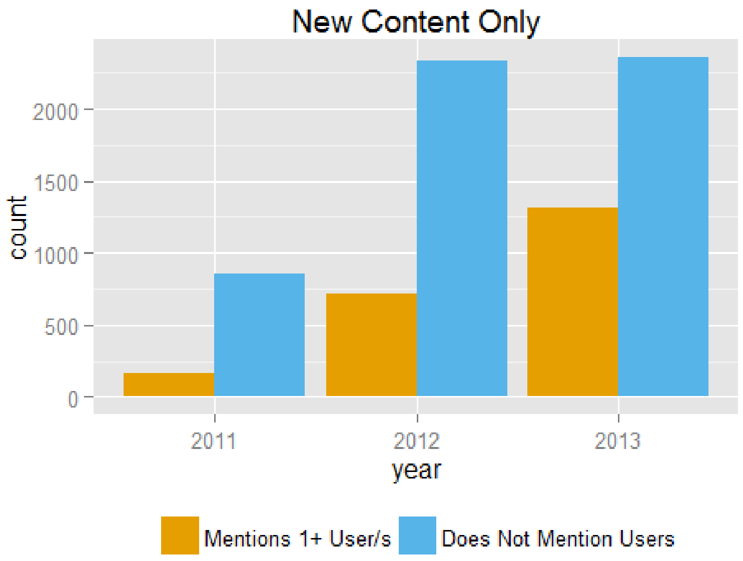


Year	Modified Content	Retweeted Content	New Content
2011	51	578	1020
2012	180	1348	3060
2013	134	2914	3671

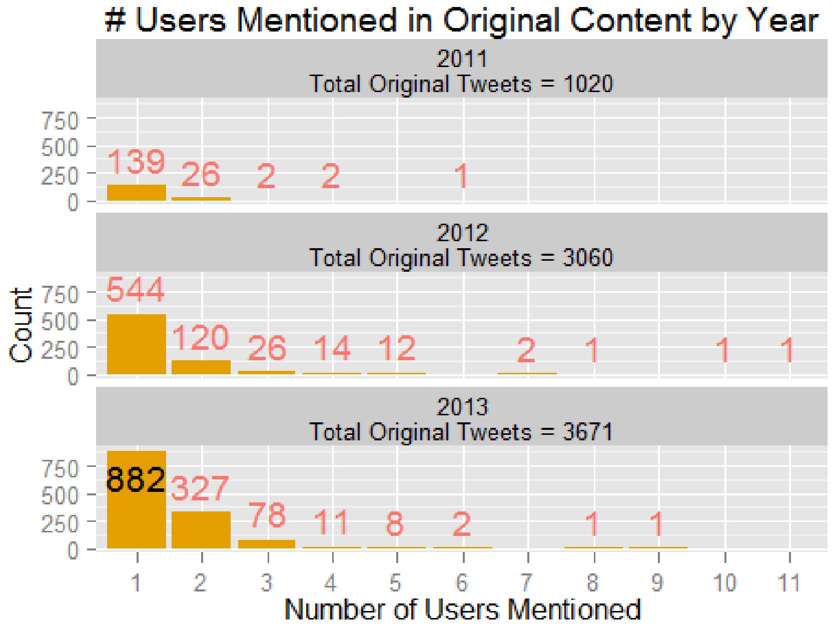
Discussions

I think one of the best things about twitter are the conversations you get into during the conference. Although we don't have a way to measure the verbal conversations and connections that occur during/following the meeting. I thought there were some text approaches we could use to identify "discussions". I took all of the tweets identified as original content in the previous analysis and looked for @[:alphanum:] in the tweet body. I counted each occurence in the tweet to get a sense of how large conversations could be.

Importantly, this analysis makes a number of assumptions. First, it is customary that if a speaker has a twitter account (and the tweeter knows their handle) the presenter is quoted using their handle. I would not really count these as conversations as the tweeter is usually not expecting a reply. I view this process a more of a stronger citation method to give credit to the presenter. Unfortunately without manual review I don't see us getting around this issue.



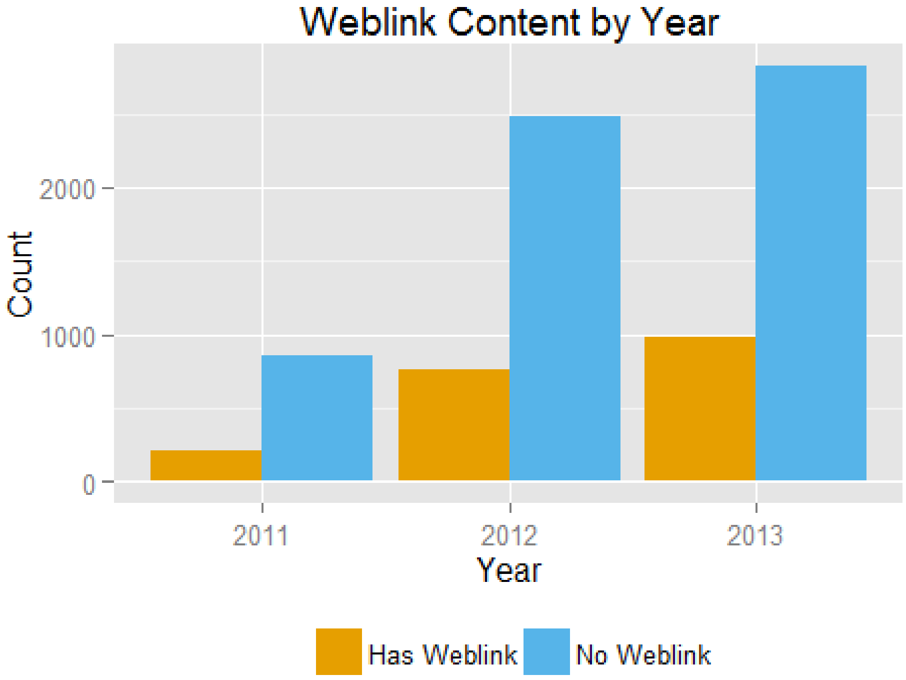
Year	Mentions 1+ User/s	Does Not Mention Users
2011	170	850
2012	721	2339
2013	1310	2361



Weblinks

The other great feature of twitter is the ability to add to the conference experience. Beyond the conversation aspect, a great asset of twitter is often getting links to new information providing immediate access to knowledge. It's a great way to share relevant papers, tools, examples, etc.

Positively, almost every link that is tweeted is wrapped with a "t.co" link from twitter. This was originally implemented to give twitter more control over malicious links. I did a search for "t.co","html", or "www" as my definition of a link. Importantly, I removed true RTs from this analysis.



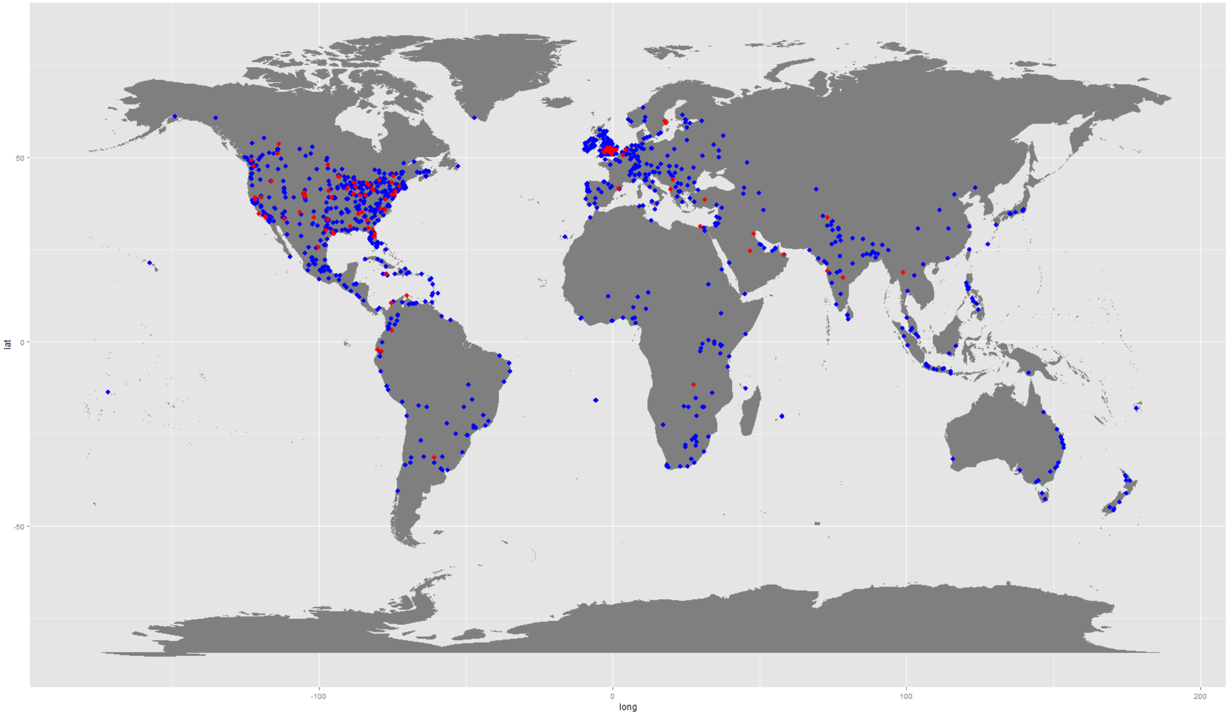
Year	Contains Weblink	Does not Contain Weblink
2011	211	860
2012	761	2479
2013	976	2829

Content Reach

Finally I wanted to get a sense of the reach of #AMIA content. Unfortunately I can't get retrospective records of tweeter's followers at the time of each symposium. However I can use their current follower list to create a map representing the location of our tweeter's and their followers today.

At the moment I am hitting API limits to get this information for all of the tweeters. However to give you a taste of what this would like... As the example, red dots are tweeters and blue dots are their followers. The method to derive location is very coarse as twitter only returns what the user provides for location. I have a script from http://biostat.jhsph.edu/~jleek/code/twitterMap.R that partiallly processes this data. However it needs to be fixed to handle more use cases.

If you think this would be valuable, we can fix the code. However, unless I hear from you I will *NOT* be continuing to tweak this analysis (either location function or pulling data from twitter).



For the this analysis I selected all unique tweeters between 2011-2013. I then queried twitter using the twitteR package. We had 1132 tweeters, but due to user's privacy settings we were only able to retrieve 1063 twitter accounts. Of those a number of them will not allow for access to their follower list (count unknown at this point). Of these, a number will not have a location for which we can derive latitude and longitude.