Paper 355-2010

Listening to the Twitter Conversation

Russell Albright, Richard Foley, and Ravi Devarajan, SAS Institute Inc., Cary, NC

ABSTRACT

Twitter, with its short texts and large number of users, zero cost of entry, and open API, has become a social media phenomenon providing an inexpensive way to connect with customers, increase brand awareness, improve product development, and gather competitive intelligence. The key, to take advantage of this information, is to understand the Twitter conversation: who are your influencers; what are they saying when they talk about you; who should you follow based on what they talk about. This paper discusses the use of text mining, visualization, and the HTTP Procedure to provide a complete understanding of the Twitter conversation.

INTRODUCTION

Twitter is one of the most popular micro blogging sites in the world. (Micro blogs are restricted to small amounts of text; for example, the Twitter limit is 140 characters. A regular blog is essentially unlimited in the amount of characters and content allowed.) Twitter has become so popular that a lexicon has been created around it:

- Tweet – the content posted to Twitter
- Tweeter – a person who has sent out the tweet
- Hashtag – a way to represent a subject the tweet is about
- Retweet – the process of forwarding someone’s tweet
- Mention – the process of referencing a person in a tweet
- Followers – people who are interested in what a person tweets about
- Friends – people who the tweeter follows to listen to their tweets

Here is an example of a tweet:

```
Tweeter: richardfoley
```

RT @sascustomer finding great information at #SGF10 thx @saspublishing go #SAS.

The abbreviation RT means this is a reposting, called a retweet, of a tweet by tweeter @sascustomer. The original post from @sascustomer is *finding great information at #SGF10 thx @saspublishing go #SAS.* The term #SGF10 is a hashtag representing a topic that the tweet is about, #SGF10 in this case refers to SAS Global Forum 2010, and #SAS represents the software company SAS. It is up to the tweeter to add tags to his tweet. When he does this, the tag becomes a form of metadata describing his tweet.

What makes Twitter so exciting to analyze is its openness. Twitter has APIs that allow access not only to the tweets of users but also to a user's profile, who a tweeter follows and who follows that tweeter. SAS provides a procedure called PROC HTTP that connects SAS to the Twitter API and downloads information or even tweets from SAS. The following is an example of code used to access the Twitter search API:

```
/*
*  This program will make a request for #SGF10 to the Twitter search API.  The API
*  will retrieve the most recent tweets, for the amount Twitter will allow, for any
*  tweets with the term #SGF10.
*  The query request for the search API is q=”%23sgf10”&ppr=1500.  %23 represents the
*  # and ppr means pagers per request.  Which means 1500 tweets per page, which is the
*  maximum number of tweets per page.  There are other options in the Twitter API,
*  such as a page that will allow you to download other tweets.
*  Put the query string into a temp file.  For the search API there are two types of
*  formats:  atom and json.
*  ATOM is an xml format and the one we will use for this program.
*  Then use PROC HTTP to make the request to the Twitter API, which will download the
*  tweets to filename twtOut. Because the data is in xml form we will use an XMLMap
*  and the xml LIBNAME engine to load the data into SAS. (The use of automap in
*  XMLMapper will easily map the xml to data sets, which will be loaded into SAS.
*
*/
   filename REQUEST temp;
   DATA _NULL_;
   FILE REQUEST;
   INPUT;
   PUT _INFILE_;
   CARDS4;
   q=%23SGF10&ppr=1500;

   filename twtOut "\\twitter\out\SASTweets.xml";
   %let twUser='USERNAME';
   %let twPass='PASSWORD';
   proc http
       in=REQUEST
       out=twtOut
       url="http://search.twitter.com/search.atom"
       method="get"
       proxyhost=PROXYHOST
       proxyport=PROXYPORT
       webusername=&twUser
       webpassword=&twPass;
   run;

   filename  SXLELIB "\\twitter\out\SASTweets.xml";
   filename  SXLEMAP'"\\twitter\out\SASTweets.map”;
   libname   SXLELIB xml xmlmap=SXLEMAP access=READONLY;

   libname tags '\\twitter\tweets\data';

   data sxlelib.tweettable;
       set tags.tweets;
```

For information about Twitter's APis, see http://apiwiki.twitter.com/.

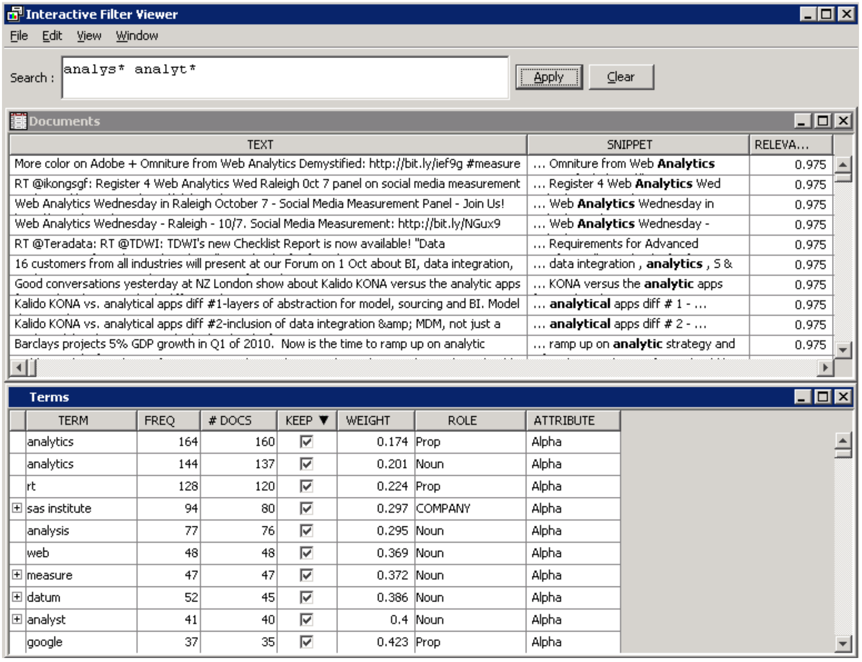
For the purposes of this paper, we scheduled a daily download of the tweets containing #SAS and #Analytics tags. We then downloaded current tweets for all the people who tweeted about #SAS or #Analytics. This allowed us to analyze other content about who was tweeting about SAS or about analytics.

EXPLORING THE TWEETS

Before proceeding in any text mining application, it is important to gain an understanding of the content that you are analyzing. This can be done by creating reports, clustering, identifying important terms and concepts, and simply searching the collection with terms of interest. In the following subsections, we show how to find specific tweeters of interest and how to report on a specific tweeter.

WHO IS TALKING ABOUT MY SPECIFIC INTERESTS?

SAS® Text Miner 4.2 includes the Teragram search engine to facilitate the interactive search of the collection that you are analyzing. In addition to standard searches, users can add meta characters to query terms in order to fine tune their search. In Figure 1, we show a wildcard character being used, but other options include requiring or excluding a term and searching for a term and all of its synonyms simultaneously.



Search : analys* analyt*

TEXT	SNIPPET	RELEVA...
More color on Adobe + Omniture from Web Analytics Demystified: http://bit.ly/ief9g #measure	... Omniture from Web **Analytics**	0.975
RT @ikongsgf: Register 4 Web Analytics Wed Raleigh Oct 7 panel on social media measurement	... Register 4 Web **Analytics** Wed	0.975
Web Analytics Wednesday in Raleigh October 7 - Social Media Measurement Panel - Join Us!	... Web **Analytics** Wednesday in	0.975
Web Analytics Wednesday - Raleigh - 10/7. Social Media Measurement: http://bit.ly/NGux9	... Web **Analytics** Wednesday -	0.975
RT @Teradata: RT @TDWI: TDWI's new Checklist Report is now available! "Data	... Requirements for Advanced	0.975
16 customers from all industries will present at our Forum on 1 Oct about BI, data integration,	... data integration , **analytics** , S &	0.975
Good conversations yesterday at NZ London show about Kalido KONA versus the analytic apps	... KONA versus the **analytic** apps	0.975
Kalido KONA vs. analytical apps diff #1-layers of abstraction for model, sourcing and BI. Model	... **analytical** apps diff # 1 - ...	0.975
Kalido KONA vs. analytical apps diff #2-inclusion of data integration &amp; MDM, not just a	... **analytical** apps diff # 2 - ...	0.975
Barclays projects 5% GDP growth in Q1 of 2010. Now is the time to ramp up on analytic	... ramp up on **analytic** strategy and	0.975

	TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
	analytics	164	160	☑	0.174	Prop	Alpha
	analytics	144	137	☑	0.201	Noun	Alpha
	rt	128	120	☑	0.224	Prop	Alpha
⊞	sas institute	94	80	☑	0.297	COMPANY	Alpha
	analysis	77	76	☑	0.295	Noun	Alpha
	web	48	48	☑	0.369	Noun	Alpha
⊞	measure	47	47	☑	0.372	Noun	Alpha
⊞	datum	52	45	☑	0.386	Noun	Alpha
⊞	analyst	41	40	☑	0.4	Noun	Alpha
	google	37	35	☑	0.423	Prop	Alpha

Figure 1. Tweet Search Results for analys* analyt*

The search subsets the data by returning all the documents in the collection that match the query and all the terms contained in those documents. Users can customize results in a SAS® Enterprise Miner™ Code node so that further analysis can be done on the subsetted documents. For example, in Figure 2 we show which tweeters were more likely to match the query. The tweeter Quantivo uses the term *analytics* in over 60% of his posts. Most of the tweeters use the term far less frequently. Quantivo might be a tweeter that you would want to pay attention to if analytics was an important keyword to you.

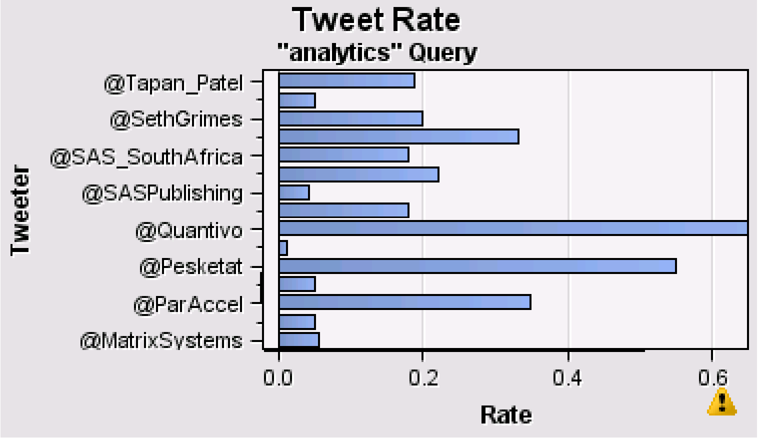


Figure 2. Tweeter’s Rate of Usage for the Term "analytics"

Both the search and the graphic on the query show the tweet rate can be automated completely with SAS, allowing you to build custom solutions around the tools. Figure 2 was created with the %EMREPORT macro.

WHAT ARE SPECIFIC TWEETERS TALKING ABOUT?

Once we identify particular tweeters, a summary of the content that they tend to tweet about is useful. In Figure 3, we show the most frequent terms (after stop lists have been applied) for four tweeters of interest. The tag clouds presented in this graphic assist us in quickly seeing what individuals are discussing. The largest terms in each cell represent the most frequently tweeted terms by the corresponding tweeter, while the color indicates the number of times a particular term is used by other tweeters. A larger term of lighter color, such as *programmer* or *statistic*, means that the term is characteristic of the tweeter. Larger terms of darker color, like *day* or *love*, indicate terms that are not only tweeted by the author but also other tweeters.

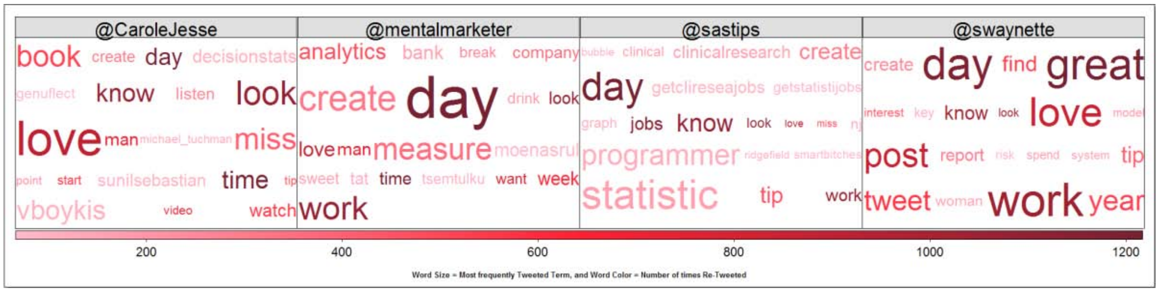


Figure 3. Tag Clouds for Four Tweeters

The tag cloud diagram is not currently available in SAS but will be available in an upcoming release.

MONITORING THE TWITTER FEED

The Twitter feed can take the pulse of public opinion about the products and customer service that a company offers. Discovering trends from Twitter allows companies to act quickly to important feedback. Occasionally, completely unexpected events can influence the discussion surrounding your business. In this section, we show you how to use the time component of tweets to help you make important decisions about customers.

CREATING THE TIME WINDOWS

SAS Text Miner can be used to separate mundane tweets from interesting tweets by monitoring incoming tweets for important changes over time. This can be done by creating time windows and processing the tweets within each time window as a single collection. Given a date stamp on the tweet data, the INTNX function allows you to normalize a date to a specific interval. The following is an example of its usage:

```
newdate=intnx(‘DAY’,datevar,0,'END');
```

The first parameter indicates the length of the interval that we want to use to create the time window. While SAS supports over 20 different date, time, or date time intervals, we use DAY for this example so that every timestamp will be mapped to the particular day it occurred. The second parameter, datevar, is the name of the variable that holds the datetime timestamp returned from the Twitter feed. The third parameter, 0, tells the function to return the normalized date without incrementing it. Finally, the END parameter indicates the point in the interval at which to normalize the data.

A second approach to creating time windows is to simply use a fixed number of observations. So, for example, every 1,000 tweets could determine a time window. While your intervals no longer correspond to the natural pattern of days, weeks, months, and so on, the fixed number of observations in each interval makes comparisons between consecutive windows easier to model.

LEARN FROM THE PAST

Over time, as you collect tweets and process time windows, various reports and visualizations can assist you in understanding the tweet space surrounding your retrieved tweets. You might learn about influential events like a conference that discussed your new electronic device, how the weather affected people’s opinion of it, or when a large department store chain offered it on sale.

In this section, we show an unsupervised method for discovering which terms are interesting by identifying the terms with dramatic changes in their usage rate from one time interval to another. The topics in the current time window can be compared to the previous time window to determine what remains constant and what changes in the distribution of the terms and topics discussed.

For each term in each time window, we compute the term frequencies for the tweets in each interval. Because a binomial test in PROC FREQ would require us to represent the data in an inefficient way, we use a Bayesian and information theoretic approach to compare the rates of change between terms in consecutive time intervals.

We model the rate at which each term occurs using a beta prior distribution. Instead of comparing the term frequencies, we compare the posterior distributions that model the rates with the Kullback-Leibler divergence. For more information about this approach and others, see the references at the end of this section.

The terms whose distributions are the farthest apart are reported as having a significant change. The method identifies sudden bursts or spikes in the data rather than long-term trends.

In Figure 4, we show a line plot of the biggest movers over the same interval described in Figure 3. The top six terms occurring more than 20 times in at least one of the compared intervals and having the largest normalized divergence scores were selected for display.

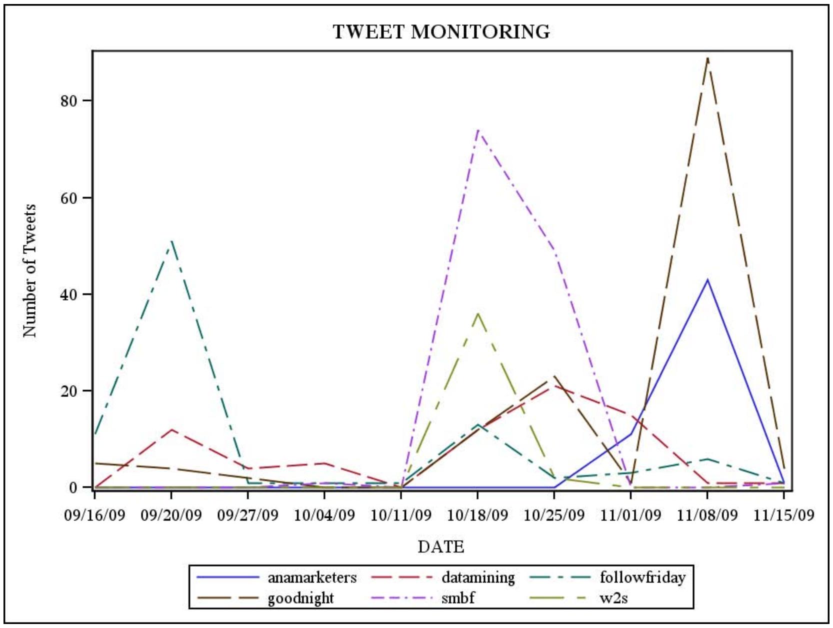


Figure 4. Biggest Movers over Consecutive Time Intervals

Most of the results shown in Figure 4 have an interpretation that can be discovered by using the search framework shown in Figure 1. For example, the spike for *goodnight* at week 11/08 can be attributed to a speech given by Dr. Jim Goodnight at the Churchill Club and tweeted about by many of the attendees. Many of the one liners delivered by Dr. Goodnight during the speech were retweeted frequently in that time period. The *smbf* spike at week 10/18/09 is a hashtag corresponding to Social Media Business Forum. In this case, the SMBF conference was held near the Cary campus and attended and tweeted about by several SAS employees.

MONITOR THE PRESENT

In addition to viewing historical trends, this same approach also allows you to be notified in near real time as new topics or trends occur. If we shorten the current window down to a subinterval, the terms that are bursting can be identified with this technique and a notification sent. For example, within hours after the first *smbf* occurrence in the 10/18/09 interval shown in Figure 4, our model could detect that the change in the rate of usage of *smbf* was significant relative to the previous week.

UNDERSTANDING INFLUENCERS

As a company, you might want to identify and track people who are particularly influential. In many cases, influence is a simple measure of the number of followers, plus other metrics that might include some factor of retweets. A very practical measure of influence is related to the amount of retweeting done by the rest of the tweeters in response to your tweet.

Figure 5 shows a visualization of frequent tweeters and retweeters and indicates who is most influential. In the graph, the nodes represent tweeters and the size of the node corresponds to the number of times the author has been retweeted. The smallest nodes in the graph are authors who only retweeted others but never had one of their tweets retweeted. The large nodes are users who were retweeted very frequently. Because displaying the entire graph is impractical, this particular plot has been subset to show only those individuals who had their tweets retweeted at least four times or who retweeted others at least four times.

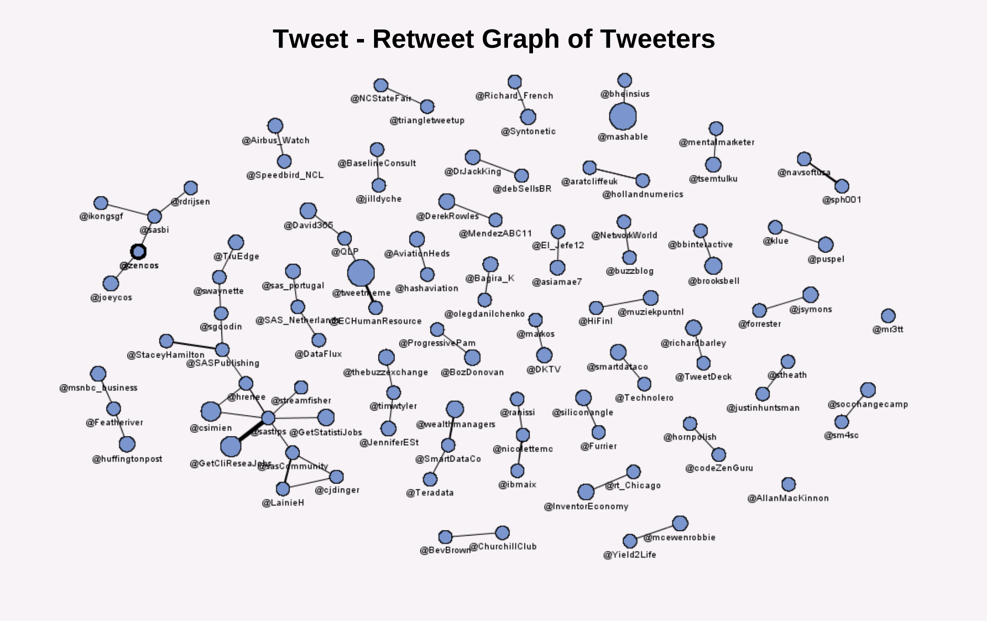
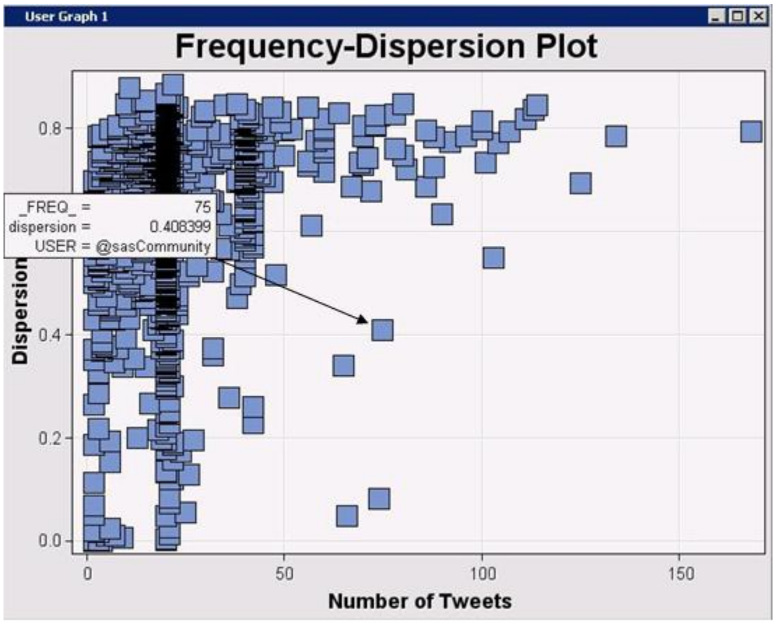


Figure 5. Most Frequently Retweeted Tweeters

Those who are retweeted the most are easily spotted in this diagram (although not all of those who retweeted that particular tweeter are displayed). The larger connected components reveal even more influence as the tweeter's tweets can be retweeted indirectly, by followers of followers.

The metrics discussed here are useful but do not include the tweeters' propensity to tweet about the topic of interest. We introduce a measure we call *dispersion*, which gives an idea of how focused a tweeter tends to be. For each tweeter, we calculate how far, on average, his tweets are from the mean of his tweets. Figure 6 displays each tweeter's dispersion score against the number of tweets he has made. Those with a higher frequency of tweets but a low dispersion suggest that the tweeter is quite focused on a topic and is not tweeting about daily random activities.



FREQ = 75
dispersion = 0.408399
USER = @sasCommunity

Figure 6. Scatterplot of Tweet Frequency by Tweet Dispersion for Tweeters

ISSUES

Twitter is a great source of information. It is amazing what can be packed into 140 characters. However, with only 140 characters, there are a lot of abbreviations, missing words, and nonstandard usages. Hashtags have no standard, so #SAS could mean SAS airlines, Sing along Saturday, or Sailors at Sea. In addition, the 140-character limit means a lot of people will embed blog posts and Web pages. This can be important information to include in an analysis. As with all textual information, topical knowledge is extremely important, maybe even more so in Twitter where the Twitter lexicon is required.

Analysis of Twitter users has shown that 5% of the Twitter universe represents 75% of the tweets. So not only are you limited in information but you could be limited in the number of users who are actually tweeting about your subject of interest. This can easily skew results, to the squeaky wheel syndrome.

It is difficult to judge what is being heard. Just because someone has 1,000 followers does not mean that the followers are constantly listening to Twitter.

CONCLUSION

Though there are issues with Twitter, there is still a wealth of information that can be gleaned. SAS provides great tools to understand Twitter, from collecting the data with SAS procedures (like PROC HTTP) to analyzing the data with SAS® Text Analytics to visualizing the data with SAS/GRAPH® software. With the right visuals, it is easy to discover trends, fads, and other relevant information that might elude you otherwise. More importantly, social media information can typically be downloaded using various APIs. These APIs tend to be Web services: SOAP based for PROC SOAP or HTTP based, such as Twitters, using PROC HTTP. These two procedures allow SAS to access the information available in social media. Furthermore, SAS Teragram's Search Engine, available in SAS Text Analytics, can also access the information found in Twitter.

Textual analytics can offer more insight than just straight term counting and association analysis. SAS Text Analytics provides the ability to analyze the Twitter conversation in its entirety, filter out the noise, and understand the topics of the conversation. Understanding individuals who focus on topics provides more clarity to what is relevant to your brand.

In order to interpret the data, you need to be able to view the data in a way that is easy to understand. New graphics being developed by SAS, and other standard SAS graphs, allow for quick and easy visualization of what is happening on the Twitter conversation.

With SAS, a complete set of products to access, analyze, and visualize the Twitter conversation is available. Better brand management and improved viral and direct marketing can occur.

REFERENCES

"Beta Distribution." *Wikipedia*. 18 February 2010. Available at http://en.wikipedia.org/wiki/Beta_distribution

He, Xiaomin,.and Wu, Shwu-Jen, "Confidence Intervals for the Binomial Proportion with Zero Frequency", PharmaSUG, March 2009.

ACKNOWLEDGMENTS

The authors would like to thank Kelly Graham for her review of drafts of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Russell Albright
SAS Campus Drive
SAS Institute Inc.
E-mail: Russell.Albright@sas.com

Richard Foley
SAS Campus Drive
SAS Institute Inc.
E-mail: Richard.Foley@sas.com

Ravi Devarajan

SAS Campus Drive
SAS Institute Inc.
E-mail: Ravi.Devarajan@sas.com