

Super crunching Twitter data

Using Twitter data to predict the success of innovations

ABSTRACT

In this paper, the predictive possibilities of Twitter will be explored. First, the characteristics of the information and data streams coming from Twitter are described. Other initiatives that crunch data from the web will also be discussed. Secondly, some 'data crunching' techniques that could be used on this data will be presented, considering a potential use case for Twitter-prediction involving the diffusion of innovations.

INTRODUCTION

Ten years ago, the internet was nothing more than a huge collection of static documents. Nowadays, the web is not just a big library, but has grown to be a popular communication medium as well. There are lots and lots of online social networks like Facebook and Hyves that enable you to keep in touch with family, colleagues and friends you don't often see (or, even worse, barely know at all) in real life. Dating sites, career networks and online auctions are pretty much accepted in our society.

People sometimes put their entire life online, by posting videos to YouTube and writing whatever they like about their day-to-day life on web logs. In 2006, a new service emerged from a brainstorm session at an internet company called Odeo (Sagolla, 2009). Twitter, as it was called, allows people to periodically post a small, 140-character message from their mobile phone (using SMS) or the internet, containing their status and whereabouts. Quickly, the service was picked up by a large number of people, including celebrities (Lance Armstrong, Oprah Winfrey and Britney Spears, to name a few) and companies. Instead of a simple 'status'-website, Twitter became a mass medium, on which small pieces of information can easily propagate. In some cases, Twitter was much earlier than regular news media in reporting important events: 'Michael Jackson' was a trending topic in Twitter hours before the national news media picked up the story of his death. Even more interesting is the fact that not all messages on Twitter come from humans: it is very easy to 'feed' the output of a website (a weather or traffic website, for instance) to Twitter.

Because there is so much real-time information on Twitter, it could be possible to use this data for other purposes than just smalltalk and entertainment. Although it may seem strange to use information that is so unstructured like this, this is exactly the point that Ian Ayres (2007) makes in his 'Super Crunchers'-book: as long as you have a lot of data, it is perfectly possible to extract useful patterns and/or predictions from it. In fact, Google is doing it already, using the terms that people search for using their search engine as only data source to predict the risk of getting the flu in a particular area (Ginsberg, Mohebbi, Patel, Brammer, Smolinski, & Brilliant, 2009).

In this paper, the predictive possibilities of Twitter will be explored. First, the characteristics of the information and data streams coming from Twitter are described. Other initiatives that crunch data from the web will also be discussed. Secondly, some 'data crunching' techniques that could be used on this data will be presented, considering a potential use case for Twitter-prediction involving the diffusion of innovations.

WHY THE TWITTER DATA COULD BE A USEFUL PREDICTOR

Why even look at using Twitter data for predicting anything, if search terms can already do a great job? Looking at what Google does, using Twitter data for predicting the risk of getting the flu is a logical next step, but it would probably only be marginally better than the existing prediction, if at all. Still, there are a number of reasons to use the Twitter data.

A first, practical reason is that the Twitter data is readily available, and can be freely used by anyone. While Google views their search term data as a very valuable company asset and won't even think about letting other parties use it, Twitter is completely on the other side (although they seem to know too that these *tweets* are a valuable asset (Williams, 2009)). Another interesting property of tweets is that they are typically much longer than the average search term (although limited to 140 characters). This means that these are more complex, but also provide more contextual information.

The second reason why the Twitter data could be interesting is the fact that Twitter aggregates many different data sources. While, in the beginning, Twitter used to be an information source itself (having users post messages like 'my cat is sick'), the Twitter of today is more like an information distribution and aggregation medium. Twitter has provided programmers with very easy-to-use interfaces that enable them to 'feed' data from one website to Twitter. For instance, an enthusiastic Twitter user is currently feeding messages from P2000, the paging system used by the police and fire department in the Netherlands, to Twitter. In addition, newspapers like the New York Times and news channel CNN are feeding their headlines to Twitter. For predictive purposes, it is clear that the 'aggregation and distribution' is not at all bad: there are not just *more* messages, but the data contained in these messages is also of higher quality.

Another aspect that makes the Twitter data interesting is the fact that Twitter is actually a social network. Twitter users can 'follow' another user, after which they automatically receive all tweets by the followed user on their personal page. Using this feature, it is easy to see what your friends are up to (and who they are following). If a particular tweet is interesting, a Twitter user can even 're-tweet' it to its own group of followers. This follows/followed-by principle creates a large network between users (Huberman, Romero, & Wu, 2008). This network data is not just useful if you want to know more about the Twitter users, it could also be used to improve the prediction.

You could, by means of network analysis, try to find out which Twitter users are important for a particular subject (i.e. are authoritative), and use this to weigh the data. Another interesting use would be to analyze how specific terms disseminate over a network of Twitter users. When the Apple iPhone was introduced, many people would post a message on Twitter when they had bought one. It would be interesting to see to what extent people are influenced by the people they follow on Twitter (do they also follow those people in real life?). It should be possible to find out which Twitter users are early adopters and/or opinion leaders in a particular context, which could be valuable information for companies that want to market a new technology.

The fourth interesting property of the Twitter data is that it (since recently) contains location information. Although Google also seems to be able to pinpoint the location of its users, it cannot do so with great precision; it relies on the IP-address of a user, and can only determine a region, sometimes even just the country. The Twitter location data is more precise: a large fraction of the tweets is posted from mobile phones like the iPhone, which have a built-in GPS-receiver. Twitter records the exact latitude/longitude information whenever the client sends it. It is possible to ask

Twitter for all the tweets that have been posted in your neighborhood (using the “*near:Eindhoven*” syntax, you can find all messages posted near the city of Eindhoven – it is also possible to specify a range in kilometers). This information could be used to find terms that are literally ‘coming at you’: it should be possible to predict a lightning storm if the Twitter data indicates that there is one a few kilometers away, and it is moving to your location. Another interesting use, related to the ‘diffusion of innovation’-example above, is to see how innovations diffuse geographically: you could identify clusters of early adopters, as well as areas where an innovation is not being adopted.

Finally, the Twitter data also potentially contains emotional data. Like in the good old chat boxes, a large number of tweets contain so-called smilies, which indicate the emotion associated with the tweet. If you want to buy a particular product and want to know whether people are happy with it, a Twitter search for ‘product name’ combined with either ‘:-)’ or ‘:-(‘ will return some interesting results. This sort of information, again, is useful with respect to innovations: firms can simply search and find the problems with their innovation, and maybe try to fix them.

THE TWITTER DATA

Before starting to use the twitter data to predict, it is important to look at the nature of the data that can be gathered from Twitter. First of all, it is necessary to check where the Twitter data actually comes from. If the user base of Twitter is only a small, non-representative group in the total population, then chances are that certain topics are not ‘twittered’ about. Secondly, we need to know a little bit more about what is in the data itself; these will be the variables that will be used for prediction later on. First, let’s look at the Twitter users.

TWITTER DEMOGRAPHICS: WHO IS TWITTERING ANYWAY?

To be able to predict anything reliably using data from Twitter, it is necessary that the data provides a representative sample of the variables you are using. For instance, when you want to predict the risk of getting the flu in a particular location by measuring the amount of *tweets* containing the word ‘flu’, there have to be Twitter users in each of the regions you are interested in. If all Twitter users are in America, your flu prediction for Europe will probably not be very reliable. If you want to predict the outcome of a parliament election, and all Twitter users are into technology, then chances are that the result of your prediction is skewed, because of the fact that the Twitter users might consider technical aspects when choosing who to vote for. These examples illustrate that in general, it is desirable that the Twitter user base be a representative sample of the world’s population.

Since Twitter does not ask many personal details upon registration, there are no definite numbers on the gender, age, location and education of Twitter users. Nevertheless, analyst firms like Nielsen have some useful demographic information. According to their data, which is from March 2009, Twitter had more than 7 million users, of which the largest group (41.7%) is between 35 and 49 years old (McGiboney, 2009). The data from internet audience analysis firm Quantcast about Twitter agrees with this, and also suggests that there are almost as much males as females on Twitter (Quantcast, 2009).

Using these numbers alone, it is hard to say whether data from Twitter will do a good job at predicting things, and if so, which things it can predict well or not. There also seems to be no demographical data for the users of Google (although QuantCast data suggests that this is a pretty good representation of the world’s population), so it is hard to compare the two. Another difference between Twitter and Google is that user effort is required on Twitter, whereas the Google search terms are entered by the user in their own interest (they just want to find something). It might be the case that there is a specific category of Twitter users that stops being active on Twitter after a

while. Also, there might be some topics that are not discussed on Twitter, while they are being searched for. People might be more hesitant to twitter about health issues than to search for them using Google, because of the fact that Twitter messages can be read by anybody.

Still, there seem to be a few categories of topics for which Twitter is perfectly suited. A quick look at the Twitter ‘trending topics’ reveals that there are a lot of tweets about technology (“*Windows 7*”, “*Google Wave*”), sports, entertainment and celebrities (“*Michael Jackson*”, “*Yankees*”, “*Halloween*”). Also interesting are the ‘Twitter hypes’ that appear every now and then (for instance, at the time of this writing there is a hype to tweet a funny new title of a movie where only one letter is different from the original title). Because a large number of Twitter users will participate in such hypes, the results of a prediction could be significantly influenced.

Nevertheless, if all the limitations and potential problems described above are taken into account, prediction should be possible and yield useful results, at least for the topics that are discussed on Twitter and for which there is a representative sample of the population. Using an algorithm that also considers the older data and/or is able to combine terms can probably mitigate the ‘hype’ - problem. In terms of linear regression, it could be more robust to count the number of times that the terms ‘*obama*’ + ‘*election*’ occurs in a tweet than the term ‘*obama*’ alone.

THE DATA PROVIDED BY TWITTER

Twitter provides a complete API (*Application Programming Interface*) and documentation website to programmers, who can easily connect their program or service to the Twitter platform. In order to predict anything from Twitter data, you would probably use the ‘stream API’ that Twitter provides. Connect to this API with a simple program, and Twitter will continuously send you a random sample of all tweets that are currently written. It is also possible to specify search terms, a location or a specific Twitter username that these messages must match.

The messages themselves are between 1 and 140 characters long and are associated with a single Twitter username that wrote it. Within the message, other Twitter users can be referred to (‘mentioned’) using the ‘@’-character (i.e. “@chris”). Also, when there is some ‘hot topic’ on Twitter, people add their tweet to the conversation by adding one or more conversation names (i.e. “#Election09”); whenever people think that some topic is going to be a hot topic, they will come up with these ‘hash tags’ by themselves¹. Because the Twitter-frontpage also lists terms without a hash tag that occur frequently, it is possible that a hash-tag is first used when a topic is already very popular.

ANALYZING THE TWEETS

In a nutshell, conversation topics on Twitter more or less magically appear and disappear. People are automatically involved, either because they are ‘following’ the messages of another Twitter user they know or has the same general interest, or because they are ‘mentioned’ in another message. It is clear that both mechanisms form an interesting structure that relates to the emergence of trends and their propagation across a social network. On the other side, these messages are complex by nature, since they are in a human language and they come in large numbers. How in the world could you analyze them?

¹ These ‘hash tags’ were invented by the Twitter users themselves and were used even before the Twitter website had any support for them. The syntax originates from the Internet Relay Chat (IRC)-system, in which each chat room had a name that started with a ‘#’, followed by an abbreviation of the topic for that chat room (i.e. ‘#dutchfootball’). In this light, you could see Twitter as an organic collection of loosely-defined chat rooms that continuously appear and disappear.

'naïve Bayesian classification' to classify twitter messages as 'ILI-related' or not. The Bayesian method is currently used in a similar way in spam filtering software. The classifier keeps a list of words and associates each of them with a probability of belonging to a certain class. When a new message comes in, the total probability for each class is calculated using Bayes' theorem. The probability values are updated by feeding the classifier a dataset of which the actual class is known. After using the Bayesian filter to filter out the ILI-related messages, a linear model could be used on the number of messages.

PREDICTING THE DIFFUSION OF INNOVATIONS

One of the central works in the field of innovation sciences is the "Diffusion of Innovations" book, by Rogers (1964). In this book, Rogers describes the mechanisms by which innovations 'diffuse' in society. The (more or less famous) graph below describes how successive groups of consumers adopt an innovation, increasing and eventually saturating the market share. As the graph below makes clear, the rate of adoption of an innovation is very low in the beginning. According to Rogers' theory, a certain critical mass is needed, after which the adoption of the innovation will be self-sustaining. This effect can (among other things) be caused by (network) externalities. This is the case when the utility of an innovation for the consumer increases with the number of people that already have adopted the innovation (such as the telephone; if more people have one, you can call more people, making the innovation more valuable to you).

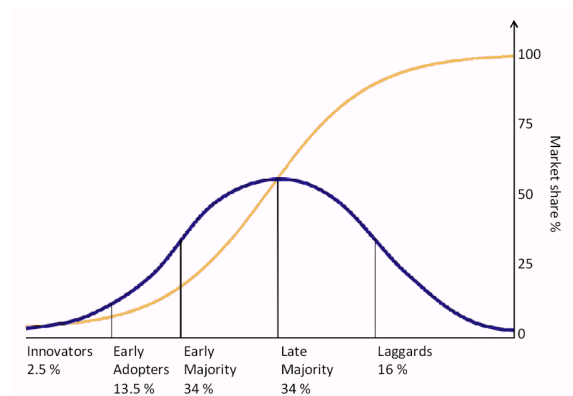


FIGURE 2 THE 'BELL CURVE' OF THE RATE OF ADOPTION OF AN INNOVATION (ROGERS, 1964). (PICTURE IS IN THE PUBLIC DOMAIN, FROM WIKIPEDIA COMMONS)

Firms that want to market an innovation will have to take these things into account to aid the adoption of their innovation. Luckily, Rogers suggests a few ways of dealing with the initially low rate of adoption (Rogers, 1964). First of all, firms could look for 'opinion leaders', influential people in a social network, and make them adopt the innovation. This is for instance the case when celebrities start using a specific gadget (like an iPhone) and is probably also the reason why Apple computers can be seen in so much movies and television series. Another strategy is to create a 'niche' in which people will adopt a certain innovation, to overcome the network externalities. If the innovation is successful in the niche, it is easier to expand it to other areas. Very often, technologies find their use in military or business environment, and are then (sometimes accidentally) 'spilled over' to a bigger market.

HOW NUMBER CRUNCHING THE TWITTER DATA COULD HELP

All these strategies however require a very thorough knowledge of the market and the potential future users of an innovation. Could number crunching the Twitter data give firms some clues? It may seem far-fetched, but if there's one thing that's often discussed on Twitter, it's new technology. As can be seen in figure 3, searching for the model name of a recently introduced phone returns quite a few messages from people who have adopted it.

One of the trending topics in the last week was 'Google Wave', a new service that allows people to easily collaborate on documents and other content. Now suppose you want to launch a similar web service and would like to know how you could 'speed up' its adoption. You could construct an adoption curve by counting the number of Twitter messages mentioning a specific similar

innovation. This may help identify phases in the diffusion process in which the rate of adoption is too low and you should try to stimulate adoption.



FIGURE 3 SEARCHING FOR THE NAME OF A NEW PHONE FROM HTC RETURNS QUITE A FEW MESSAGES FROM PEOPLE WHO HAVE ADOPTED THIS PRODUCT.

If you know who the early adopters are and where they come from, you could target your marketing to this group of users. If people from Eindhoven are more likely to adopt a web service like Wave than people in other parts of the Netherlands, well, start in Eindhoven. The Twitter data could help you identify these locations by looking at the amount of Twitter messages of an older, similar innovation and clustering them by location.

Another potentially interesting use of the Twitter data is to find out how much your innovation will rely on word-of-mouth advertising. This is pretty easy using the Twitter data, because of the followers-network and 're-tweeting'. If someone tweets about an innovation, and then a follower starts tweeting about it too, this might mean that this user is 'following' the innovation as well. In social network terms, this is called the amount of *exposure* (the number of adopters in the personal network of a person) (Valente, 1996). If you know how much the adoption of your innovation relies on word-of-mouth advertising, you know how much you should focus on that in the marketing of your innovation. Network analysis could also be used to find out who the opinion leaders are (Abdel-Ghany, 2009); firms could then approach these opinion leaders and, for instance, give them free access to the service to try it out.

HOW TO ACTUALLY DO THIS

A slight problem is how you could use Twitter data for your innovation if people are not (yet) talking about it. A possible solution is to look at similar innovations: if you are launching a web service, look at tweets that contain the names of existing, similar web services (for instance 'Google Wave'). For Google, this is even easier, since they could search for messages about some of their other web services (like Gmail). The rest of the analysis can be done using the methods described earlier: either just count the number of occurrences of a particular term (with respect to the total

number of messages) or use a classifier first to determine whether a message is related to your innovation or not.

To analyze the locations of the adopters, you will have to save all the tweet locations, and run some kind of clustering algorithm. After all, a model that only spits out an 'average longitude/latitude'-number is not very useful. Another possibility is to cluster the messages based on an existing clustering (such as countries, states, zip codes) that can be distilled from the latitude/longitude data of the Twitter messages. Then, for each region, you could calculate the average time and see which region was earliest in tweeting about a particular innovation. If you have enough data, you might even be able to classify regions into the five consumer groups Rogers describes in his theory.

All these things could be combined into one big model. The independent variables would be the location (or region/cluster) of the user, possibly the age and sex of the author and whether the author is a 'follower' of other people that have already posted messages about the innovation. The dependent variable would be the chance that the user is going to adopt the innovation, or maybe a classification into one of the customer groups in Rogers' theory. The model is then estimated with Twitter messages containing terms that indicate a similar innovation. While you will not use the model to try to predict the chance of adoption for each Twitter user, the estimated coefficient values (both for main and interaction effects) will be interesting.

ADOPTION OF THE MODEL

Firms may use Twitter as a tool to determine characteristics of the diffusion of an innovation by looking at historical Twitter data about a similar innovation. Not only can the amount of messages about a particular innovation tell something about the rate of adoption, there is also interesting metadata to consider. The location information that Twitter provides can be used to identify innovative regions, whereas the social network data can be used to identify opinion leaders and the way in which innovations diffuse. Although the conclusions from the Twitter data may not always be generalizable and sometimes limited by the topics discussed on Twitter, it is probably still cheaper and easier to do than full market research, and may provide insights that such research cannot.

Twitter data may also be used as empirical data, to be used in conjunction with theories on the diffusion of innovations. Twitter makes it easy to obtain data on the adoption of innovations that is otherwise difficult to obtain. Because Twitter also provides social network data, social networking theories may also be used and tested on the data and combined with theories on the diffusion of innovations.

The usage of Twitter data will probably not have any big societal impact; the Google Flu Trends initiative did not really receive any criticism. It may however be the case that people will behave differently on Twitter when they know that their messages are being used for predictive purposes. Anyone using Twitter data for prediction should realize that the data is easily influenced, and that it should always be used in conjunction with other data if reliable results are required.

CONCLUSION

Although predicting the diffusion of innovations using Twitter data can be done using relatively simple methods, there are also some limitations. As we have seen, the Twitter user group is not a very good representation of the total population. Second, not all topics are discussed on Twitter. A third issue is that models based on Twitter data have to be robust enough to deal with Twitter hypes and possibly deliberate manipulation. For instance, George Hotz, a programmer that created a software program to 'unlock' the Apple iPhone, asked people on his website to start twittering about this new software. Hotz would only release the software when his software was the number one trending topic on Twitter, which also happened, as many iPhone users were waiting for such a tool (Hotz, 2009).

Nevertheless, using the Twitter data has some advantages over using, say, the Google search term data (which may also be subject to 'hypes' as describes above). The messages from Twitter have more meta-data, like location and social network information of the author, which can be valuable sources of information. In analyzing the diffusion of innovations, the Twitter data is a unique source of information that provides a broad, overall view of adoption over time and space. In addition, Twitter aggregates messages from many other websites and services. It is to be expected that in the future, more and more services will be 'connected'. Also, users will have more ways to put their messages on Twitter. What goes for super crunching in general is also applicable to prediction using Twitter: the more data, the better.

BIBLIOGRAPHY

- Abdel-Ghany, M. M. (2009). Social Network Analysis of the Diffusion of Innovations. *Ekonomika ir vadyba: aktualijos ir perspektyvos*, 2 (11), 270-272.
- Ayres, I. (2007). *Super Crunchers*. New York: Bantam Dell.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature* (457), 1012-1014.
- Hotz, G. (2009, November 3). *sn0wday*. Retrieved November 7, 2009, from On the iPhone: <http://iphoneitag.blogspot.com/2009/11/sn0wday.html>
- Huberman, B. A., Romero, D. M., & Wu, F. (2008). *Social networks that matter: Twitter under the microscope*. HP Laboratories, Social Computing Lab, Palo Alto, CA 94304.
- McGiboney, M. (2009, March 18). *Twitter's Tweet Smell Of Success*. Retrieved October 30, 2009, from Nielsen News: http://blog.nielsen.com/nielsenwire/online_mobile/twitters-tweet-smell-of-success
- Quantcast. (2009, October 30). *QuantCast Audience Profile for Twitter.com*. Retrieved October 30, 2009, from Quantcast.com: <http://www.quantcast.com/twitter.com>
- Rogers, E. M. (1964). *Diffusion of Innovations*. Glencoe: Free Press.
- Sagolla, D. (2009). How Twitter was born. In D. Sagolla, *140 Characters: A style guide for the short form*. Wiley, John & Sons, Incorporated.
- Valente, T. W. (1996). Social network thresholds in the diffusion of innovations. *Social Networks* (18), 69-89.
- Williams, D. (2009, June 15). *Tweets are an asset*. Retrieved October 27, 2009, from Twitter API Blog: <http://apiblog.twitter.com/tweets-are-an-asset>