The twittering machine

Miranda Mowbray

HP Laboratories
HPL-2010-54

Keyword(s):
social information systems, twitter

Abstract:
This paper is a study of the use of Twitter by automated agents, based on data sampled in July-September 2009. It discusses the dramatic rise in rapidly-tweeting automated Twitter accounts beginning in late June 2009; some surprising behaviour by automated Twitter profiles that make direct use of Twitter's API; and techniques used for automated spamming on Twitter. Ideas are suggested for ways in which Twitter might defend against some common types of automated Twitter spam. The paper ends by outlining some general conclusions for designers of social information systems.

External Posting Date: April 21, 2010 [Fulltext] Approved for External Publication
Internal Posting Date: April 21, 2010 [Fulltext]

Published and presented at WEBIST 2010 - Sixth International Conference on Web Information Systems and Technologies, Valencia, Spain, April 7-10, 2010.

THE TWITTERING MACHINE

Miranda Mowbray
HP Labs Bristol, Long Down Avenue, Stoke Gifford, Bristol BS34 8QZ, United Kingdom
miranda.mowbray @ hp.com

This paper was presented at WEBIST 2010, 6th International Conference on Web Information Systems and Technologies, 7-10 April 2010, Valencia, Spain. http://www.webist.org.

Abstract: This paper is a study of the use of Twitter by automated agents, based on data sampled in July-September 2009. It discusses the dramatic rise in rapidly-tweeting automated Twitter accounts beginning in late June 2009; some surprising behaviour by automated Twitter profiles that make direct use of Twitter's API; and techniques used for automated spamming on Twitter. Ideas are suggested for ways in which Twitter might defend against some common types of automated Twitter spam. The paper ends by outlining some general conclusions for designers of social information systems.

1 INTRODUCTION

Twitter (http://twitter.com) is a web-based microblogging service that enables users to publish messages of up to 140 characters in length. These messages are known as *tweets*. Twitter first opened to the public in October 2006. By August 2009 there were an estimated 66 million Twitter users worldwide (Solis, 2009). In a BBC interview, Twitter's co-founder Ev Harris said that the service was "*about humans connecting with each other and often in ways that they couldn't otherwise*" (BBC, 2009). But what if one of the connecting parties is not a human, but a machine?

The title of this paper is taken from a picture by Paul Klee (Klee, 1922) which depicts a mechanical contraption containing some quirky-looking birds whose legs are attached to a long wire with a handle at one end. Turning the handle presumably causes the birds to twitter. The twittering machines considered in this paper are Twitter accounts used by automated agents, which may generate and publish large quantities of tweets or carry out other sophisticated uses of the Twitter service automatically, with little effort or attention from the account's human owner. As this paper will show, in the second half of 2009 there was a striking increase in the number of these twittering machines.

This paper is a study of automated Twitter use, based on data sampled in July-September 2009. Unlike previous studies which examined the behaviour of very large numbers of Twitter accounts, this paper considers much smaller samples but examines the behaviour of some particular types of twittering machines in detail.

The contributions of this paper include data on the sudden rise of rapidly-tweeting automated accounts beginning in late June 2009, and some ideas for how Twitter might defend against common types of automated Twitter spam.

The structure of the rest of the paper is as follows. Section 2 describes some related work. Section 3 shows the rise of rapid twittering machines, and Section 4 gives some data about the behaviour of automated users that access Twitter directly via the API (Application Programming Interface). Section 5 is about automated spam on Twitter, and possible ways to reduce it. The final section discusses a few lessons for social information systems in general.

2 RELATED WORK

There have been several studies of Twitter use based on samples of large numbers of accounts. Twitter users can choose to *follow* accounts belonging to other users. They receive the tweets published by their followees in real time. (Twitter's web site refers to "*friends*" instead of followees, but this seems an inappropriate term when one of the accounts may be automated.) Java, Song, Finin and Tseng (2007) showed that follower and followee counts are correlated and obey an approximate

power law. They detected a few automated accounts. Krishnamurthy et al. found that users who had published many tweets were more likely to have follower/followee ratios close to 1. Huberman, Romero and Wu (2009) found that most users address at most one tweet to 90% of their followees. The HubSpot study (Zarella, 2009) of 4.5 million users collected over 9 months to June 2009 found that a majority of Twitter accounts have never published a tweet; a majority follow no-one; and a majority are followed by no-one. Heil and Piskorski (2009) found the median number of tweets to be 1 rather than 0. In their sample, the top 10% most prolific users produced 90% of all tweets. Cheng and Evans' study (2009) of 11.5 million twitter accounts sampled in January-May 2009 found that 24% of all tweets were made by automated accounts posting over 150 tweets a day. Their study includes data on the top 5% of twitter accounts by number of tweets, many of which are machines, and also on social media marketers, identified by keywords contained in their profile descriptions. The top 5% accounted for 75% of all activity. 35% of the social media marketers tweeted at least once a day, compared to 15% of all accounts.

However none of these papers publish separate data on accounts that make direct use of the API or on reported spammers.

There are many blog postings and newspaper articles reporting or discussing particular instances of Twitter spam, but there appear to be few published overviews of Twitter spamming methods. (Metablocks.com, 2009) briefly describes some of the methods discussed in this paper.

3 THE RISE OF THE RAPID TWITTERING MACHINES

Twitter users have the option to make their tweets visible only to their followers, but the default is that all tweets are published in the *public timeline*, a list of recent tweets visible to all. To collect data for Figures 1-5, 7060 tweets were randomly sampled from the public timeline between July 13 and September 7 2009. These tweets were published by 6932 different Twitter accounts, of which 6729 had been created at least a day before the sampled tweet appeared.

Figures 1 and 2 show data for these 6729 accounts. In Figure 1, each point represents one such account. The x-coordinate is the number of days after October 1 2006 that the account was created. The y-coordinate is the average number of tweets per day published by that account since it was created. Human twitter users unassisted by specialized software are very unlikely to produce as many as 100 tweets a day. It can be seen that although there were a few twittering machines early on, in late June 2009 there was a sudden rise in the number of new Twitter accounts with rapid rates, and the arrival of new rapidly-twittering accounts continued until the end of the sampling period.

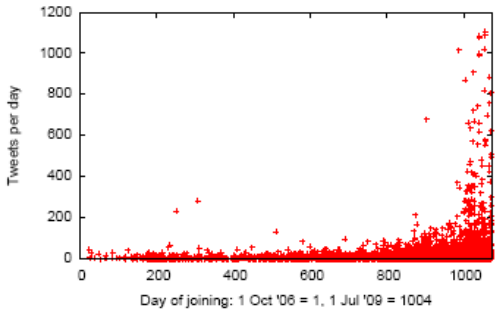

Figure 1: Tweets per day.

The more tweets an account published during the sampling period, the more likely it is to have been sampled (and an account that published no tweets during this period would definitely not have been sampled.) Therefore, the percentage of the sampled accounts that tweet more than 100 times a day will be larger than the percentage of all accounts (sampled and unsampled) that tweet at such a rate. However, it is reasonable to assume that growth in the percentage for sampled accounts reflects growth in the percentage for all accounts. Figure 2 shows that the percentage of the sampled accounts that publish more than 100 tweets a day increased from less than 1% for sampled accounts created in the first few months of 2009, to over 19% for sampled accounts created in September 2009.

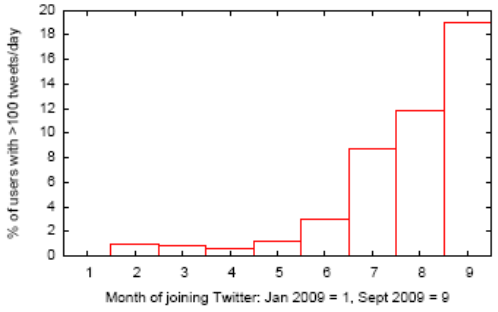

Figure 2: Sampled accounts publishing over 100 tweets a day, as a percentage of all sampled accounts created in a particular month in 2009.

Less than 0.3% of the sampled accounts that were created before 2009 had a tweet rate of over a hundred tweets a day.

In addition to the publication of the Twitter API handbook in April 2009, a probable factor in this rise is mainstream media publicity on the use of Twitter for marketing. Books about Twitter published in 2009 include *Twitter Marketing Tips* (Brooks, 2009) and *Dominate your Market with Twitter* (Smith and Llinares, 2009). Most of these rapidly-twittering accounts appear to be marketing machines. However not all of them are. The rapidly-tweeting sampled accounts include a clock that regularly tweets the time, a password generator, and a radio station that tweets its playlist.

4 API USER DATA

Twitter can be accessed via the web site, or via software packages aimed to make Twitter easier to use such as the popular TweetDeck, or twittering machines can access Twitter directly via the API. The 6932 sampled Twitter accounts mentioned in section 4 included 419 using the API. This section briefly gives some data on the follower and followee counts for these twittering machines.

Twittering machines are unable to understand other users' tweets. So one would expect the API-using accounts not to follow any other users. Surprisingly, 317 of them – about 76% of all API-using accounts – follow at least one other user, and some follow over 5,000. The question is why they should do so. Some may automatically follow their followers, as a symbolic courtesy, or to enable their followers to send them commands in *direct messages* (DMs), which are private messages sent directly from one Twitter account to another account following it. Some useful twittering machines that interact with their followers through DMs are described by Poland (2007). However, only 12 of these accounts have equal numbers of followees and followers. A likely answer to the question is that many of the API-using accounts are *follow-spammers*. Follow-spam is described in Section 5.3.

The HubSpot paper shows that there is a spike at around 2000 followees in the logscale graph of the distribution of followee counts for their data, and says that this is because "*Twitter limits users to only following a maximum of 2000 users until they have more than 2000 followers*" (Zarella, 2009, p.4). In fact, a study of the sampled data reveals that this is not quite correct. Twitter has been reluctant to reveal the precise rule, so it will not be revealed here; however, the quoted description is close. The followee count graph for the 419 API-using sampled accounts has the same feature, more prominently than in the graph for all 6932 sampled accounts.

Just over half (56%) of the 317 API-using accounts with at least one followee have at least as many followers as followees, and a few have 10 times or more followers than followees. This demonstrates that these machines are able to take advantage of their automation to gain followers by methods other than simply relying on followees to follow them back.

5 TWITTER SPAM

This section is a description of some common automated spamming techniques used on Twitter, with some suggestions of ways Twitter might defend against them.

Twitter has a spam problem. As well as causing annoyance to users, spam diminishes the quality of Twitter's search results and the potential value of the business. It is not clear whether this problem can be overcome: it is possible that by the time this paper appears, Twitter's popularity will have collapsed as a result of an ever-increasing spam burden. Not all current spamming techniques are reported here, and spamming behaviour will probably continue to evolve their behaviour over time and in response to any defences, so this section is not a comprehensive guide to Twitter spamming techniques.

5.1 Data collection

The collection of spammers discussed in this section was made by collecting the set of all the accounts reported by users to Twitter's special spam-reporting account @spam during a 24-hour period. The profiles of these reported spammers were obtained from the Twitter site. Each profile and spam report was checked, to eliminate from the collection any accounts that were reported for unwanted behaviour other than spamming or were the subject of obviously-false spamming reports (such as revenge reports, in which user 2 reports user 1 shortly after user 1 has reported user 2). This process was repeated during six other 24-hour periods. After the eliminations, a total of 1257 distinct reported spammers remained. Almost all of these appeared to be automated accounts. 261 were clones, clearly produced by just one spamming organization or individual. (There were other families of clones in the sample, but this was by far the largest).

Sections 5.2 to 5.5 describe some automated spamming methods used by these reported spammers.

5.2 Mention-spam

It is standard practice on Twitter to include a user's screen name in a tweet if it is addressed to the user, or is a copy of a tweet originally sent by them. Such tweets are displayed to the user in real time. Mention-spammers include users' names in their tweets so as to push the tweet to the mentioned user.

Some twittering machines that practice mention-spam target their messages by "replying" to the authors of tweets in the public timeline, and/or selecting which users to mention based on the presence of key words in the users' tweets.

The ability of users to contact other users not (yet) in their social network by mentioning the appropriate screen name in a tweet is a useful feature of Twitter. However, non-spamming users typically use this feature to contact many fewer users than than mention-spammers do.

One way for Twitter to reduce the amount of mention-spam is to limit the number of different screen names that a user could mention over a set time interval. Tweets that would send the count over the limit would not be published. Names of accounts following the user and accounts that recently mentioned the user in a tweet would not be included in the count.

This requirement could be weakened if it turns out to cause problems for some non-spamming uses. For example users might be able to opt to allow all other users (or some individually specified users) to address tweets to them without this increasing the user's count. It would still be possible to mention large numbers of different screen names, but doing so would incur a cost in time or effort, and this should decrease the activity of mention-spammers.

5.3 Follow-spam

Several techniques for automated spam on Twitter require the twittering machine to follow as many users as possible. This activity causes annoyance to Twitter users, who find themselves having continually to deal with new followers that are not actually interested in the user's tweets, only in sending spam.

One type of follow-spam is rapid-reaction reply-spam. Followers receive in real time all the tweets published by their followees, even if the followees hide their tweets from the general public. Rapid-reaction reply-spammers can send immediate "replies" to their followees' tweets. For instance, a followee who publishes a tweet containing a trigger word may immediately receive a "reply" that is an advertisement relating to the trigger word. The spammer can be sure that the followee is currently reading Twitter messages.

Another reason that spammers want to follow other accounts – particularly very popular accounts - was pointed out by Joel Mackey, who is @webaddict on Twitter (Mackey, 2009): the profile for an account displays a list of links to all the accounts following it, and readers of the first account may click on these links. The most recent follower is displayed at the top of the list, which gives spammers an incentive to repeatedly follow, unfollow and follow the same account again. Twitter might reduce this repeated follow/unfollow behaviour by using a different list ordering.

However, probably the most common reason for spamming software to follow Twitter users is to entice the followees themselves to follow the twittering machine back. Followers receive all their followees' tweets, and moreover followers can be sent DMs. (As mentioned earlier, these are private direct messages.) Since these do not appear on the public timeline it is hard to do crowd-policing of direct message spam, and also it creates a tension between the degree of privacy of DMs and Twitter's ability to detect and foil spammers.

How likely are followees to follow back? In the early days of Twitter it was considered impolite not to follow back, and a notification that an account has started following you is called a *friend request*. However, follow-spam may have made users more wary. In an experiment by Catalin and Carmen Cosoi, about 5 in 10 Twitter users followed back (Cosoi & Cosoi, 2009). The account used in the experiment for following users had a non-default profile image and more than a few tweets, because the experimenters had previously discovered that follow back rates were very low for accounts with a default image or not many tweets.

Among the sampled spammers, the follower/followee ratios were low for the 447 sampled spammers that had tweeted less than 5 times (over their entire Twitter history) and had at least one followee: the median ratio for these spammers was just 0.03. The median ratio for the 606 spammers who had tweeted at least 5 times and had at least one followee was 0.42. This is close to 5 out of 10, but smaller, possibly as a result of followers unfollowing the spammer before they were sampled.

Twitter's anti-spam mechanism effectively limits the number of a spammer's followers to around 2000 per account, unless users really are interested enough in the spammer's messages to follow back with a high probability. One method by which follow-spammers overcome this limit is to create multiple accounts. Another is to *churn* followees. To do this, the twittering machine follows a set of users, waits for some to follow back and spams them, unfollows those that have not followed back and those that have received a certain amount of spam, and follows other users to replace the unfollowed ones. Thus the machine never has enough followees at one time to trigger the mechanism, but over time it has a large total number of followees, and a proportional number of followers.

The software package TweeAdder was advertised to Twitter marketers with the slogan "*Auto-Follow, Auto-Unfollow, Auto-Tweet & DM It and Forget It*!"

Here is a suggestion to reduce the amount of follow-spam: if account 1 is not following account 2, and account 2 wishes to follow account 1, require account 2 to solve a CAPTCHA first. Thus, twittering machines could not follow accounts other than their followers without human assistance. Unlike Twitter's current mechanism this would restrain follow-spammers who create multiple accounts or churn their followers. This requirement could be weakened if it turns out to cause problems for some non-spamming automated uses. Users might be allowed to opt in to waive this requirement for accounts that wished to follow them. Marketers might be able to purchase a waiver from Twitter of the requirement for their accounts to solve up to a certain number of CAPTCHAs.

Twitter requires a CAPTCHA solution each time a new account is created, and some spammers are prepared to solve hundreds of CAPTCHAs so as to obtain a large number of accounts. However, if a solution was required every time a non-following user was followed by one of their accounts this would greatly increase the amount of CAPTCHAs required to carry out follow-spam, and hence would increase the amount of resources (human effort or money to pay someone else to solve them) required for this. There does not appear to be an easily-available automated method of CAPTCHA-solving at the moment: this is one of the tasks that online criminal organizations actually pay people to carry out.

5.4 Trend Abuse

Twitter's search page prominently displays a list of current *trends*. These are the 20 words or phrases appearing most often in recent tweets. (Common words such as "of" and "the" are excluded). Clicking on one of the trends displays the most recent tweets in the public timeline that contain the trend. Twitter's API also provides a list of the top trends per hour.

Trends originating from spam, including tweets sent by Facebook games, accounted for 3% of all the top trends per hour during 8 days in July 2009. In addition to spam-originated trends there is the larger issue of spammers whose twittering machines are programmed to jump on trend bandwagons, including current trends in their tweets. Since the selection of the trend is automatic, the tweet is unlikely to contribute to the discussion of the trend. Such trend-abusing tweets will be seen by users who click on the appropriate trend in the list, and by users who search for the trend (which is after all currently popular, and so likely to be a common search term). These machines have the effect that a search for the most recent messages on a trendy topic will typically reveal a large number of tweets from spammers, even if the topic was not originated by spammers.

Twitter might be able to address trend abuse by introducing a reputation measure for Twitter accounts, and only displaying tweets in search results that were published by accounts with high enough reputation. However the reputation system would need to be designed with care.

There is a special case of trend abuse that can be addressed more easily: multiple-trend spam. Some spammers include more than one recent trend in their tweets, (or even include all 20 in a single tweet!) to multiply their chances of their tweets being read. Of course, this also multiplies the amount of annoyance they cause to users who are interested in reading tweets that are actually about a trend. It should not be difficult for Twitter to reduce multiple-trend spam by detecting tweets containing multiple trends, and excluding these tweets from the results shown from a search on a trend.

5.5 Fake Retweets

Another spamming technique is to abuse Twitter's "retweet" convention to make it appear that a spammer's tweet was originally published by another user. Twitter's search capabilities could be used to match retweets with originals, and thus to detect and ban accounts publishing fake retweets.

6 LESSONS FOR SOCIAL INFORMATION SYSTEMS

There are parallels to the problems of automated Twitter use in other web-based social information systems that were designed for human-to-human communication, but have proved vulnerable to unwanted machine-to-human communication.

Forbidding automated use is not the solution. Twitter's opening up of their API to the public has resulted in some useful and entertaining twittering machines, and is likely to stimulate the development of positive new Twitter uses.

The solution is rather to create technical limits to the automated use of the system so as to allow non-automated use to flourish. This may be done by increasing the cost (in money, time or human effort) of performing particular automated behaviours. The behaviours to target are ones that decrease the usefulness of the system for non-automated users, without being essential for legitimate marketing that may provide revenue for the information system. The suggestions in Section 5 propose limits of this type.

Another observation is that to ensure that marketers do not make a nuisance of themselves in a social information network, it is not sufficient that marketing messages are opt-in only. For example, consider Twitter spammers that only send spam to their followees, using DMs. They only spam users who have opted to follow one of their accounts. However, such spammers an incentive to catch the attention of users and try to persuade them to opt in, for example by following many users, publishing automated "reply" tweets, or abusing trend words. This attention-catching behaviour can itself be an annoyance, even to users who never opt in. Designers of social information systems with opt-in marketing should try to ensure that it is not easy for marketers to use automation to do a large amount of attention-catching at a small (or zero) cost.

Access control mechanisms may help to address these problems for information systems that are not open to the public. They are less useful for a public system such as Twitter, although it could be argued that some of the limits on Twitter use suggested in Section 5 are access control rules for particular Twitter capabilities. Content validation may also help protect against some kinds of automated misbehaviour. For example, if it is possible to have a service within a social information system that could check that shortened URLs published in the system do not lead to known phishing or malware-spreading sites, this could be rather useful.

ACKNOWLEDGEMENTS

Thanks to Martin Arlitt and Phillippa Gill for their assitance, and to the New York MOMA for their permission to use the Klee picture in my presentation.

REFERENCES

BBC, 2009. *BBC Newsnight* television interview with Ev Harris, 5 August 2009.

Brooks, D. (ed), 2009, *Twitter marketing tips*, Emereo Pty Ltd.

Cheng, A., Evans, M., 2009. Inside Twitter: an in-depth look inside the Twitter world. Sysomos white paper, June 2009. http://www.sysomos.com/insidetwitter/

Cosoi, A.C., Cosoi, M., 2009. A fractal approach to social network spam detection. In *VB2009, Virus Bulletin Conference*. Virus Bulletin Ltd.

Heil, B., Piskorski, M., 2009. New Twitter Research: Men follow men and nobody tweets. Blog posting, 1 June 2009. http://blogs.harvardbusiness.org/cs/2009/06/new_twitter_research_men_follo.html

Huberman, B., Romero, D. M., Wu, F., 2009. Social networks that matter: Twitter under the microscope. *First Monday* 14 (1-5), January 2009.

Java, A., Song, X., Finin, T., Tseng, B., 2007. Why we Twitter: understanding microblogging usage and communities. In *Joint 9th WEBKDD and SNA-KDD workshop*. ACM Press.

Klee, P., 1922. The Twittering Machine [*Die Zwitscher-Maschine*], in the collection of the Museum of Modern Art, New York, http://www.moma.org . Digital image: http://www.moma.org/explore/collection/provenance/items/564.39.html

Krishnamurthy, B., Gill, P., Arlitt, M., 2008. A few chirps about Twitter. In *1st Workshop on Online Social Networks.* ACM Press.

Mackey, J., 2009. The mystery behind FOLLOW and UNFOLLOW on Twitter revealed. *Open Press Wire* article, 8 March 2009, http://openpresswire.com/internet/the-mystery-behind-follow-and-unfollow-on-twitter-revealed/

Metablocks.com, 2009. Beware of Twitter spam – an overview and guide. Blog posting, 29 June 2009. http://www.metablocks.com/blog/2009/06/29/beware-of-twitter-spam-an-overview-and-guide/

Poland, S., 2007. First Twitter bots launched: sports teams, weather, stock quotes. Blog posting, 4 April 2007. http://blog.stevepoland.com/first-twitter-bots-launched-sports-teams-weather-stock-quotes/

Solis, B., 2009. Revealing the People Defining Social Networks. Blog posting on *PR 2.0*, 1 October 2009. http://www.briansolis.com/2009/10/revealing-the-people-defining-social-networks/

Smith, J., Linares, J., 2009. *Dominate your market with Twitter*, Infinite Ideas Ltd.

Zarella, D., 2009. State of the Twittersphere, June 2009. Hubspot.com white paper, http://blog.hubspot.com/Portals/249/sotwitter09.pdf