The Role of Microblogging in OSS Knowledge Management

Jonathan Lewis

Hitotsubashi University, 2-1 Naka, Kunitachi-shi, 186-8601 Tokyo, Japan
jonathan_lewis@mac.com
http://www.lewis.soc.hit-u.ac.jp

Abstract. Given that microblogging has been shown to play a valuable role in knowledge management within companies, it is useful to understand how it is being used in relation to OSS. This project studies tweets related to 12 open source projects and keywords, ranging from web content management systems (CMSes) to general office applications. It found considerable differences in the content and exchange of tweets, especially between specialist products such as CMSes and office suites such as OpenOffice. Tweets concerning the more specialist projects tended to provide information rather than updates on the user's current status. We found a high proportion of event-driven traffic for some CMS projects, and a lower proportion for the office products and groups of projects.

Keywords: microblogging, twitter, knowledge management.

1 Introduction

In any knowledge-intensive project or organization, informal communication is vital to the timely spread of information and ideas. Considerable research has been undertaken on the role of microblogging services such as Twitter in knowledge management within enterprises [1–5]. Many OSS developers and users also use Twitter, but little attention has been paid to the role played by microblogging in exchanging and diffusing knowledge in OSS projects. This study of OSS-related microblogging explores what kind of information is being exchanged on Twitter regarding open source software, and the different ways in which Twitter is being used. In order to answer these questions, a study was made of statuses (Tweets) related to 12 OSS projects, using a taxonomy adapted from previous research on intra-enterprise microblogging. The projects selected had an emphasis on, but were not restricted to, web content management. Table 1 gives an overview of the projects and the numbers of statuses collected, sampled and analyzed.

This paper is organized as follows: Section 2 discusses the selection of projects, presents statistics about the data used, and discusses the challenges in collecting and cleaning Twitter data. Section 3 contains the findings and analysis. Section 4 discusses threats to validity and topics for further research.

E. Petrinja et al. (Eds.): OSS 2013, IFIP AICT 404, pp. 140–152, 2013.

Table 1. Overview of Twitter Statuses Retrieved and Sampled, 7 May-26 Dec 2012

Project/Keyword	Project type (language)	Number of statuses collected	Number tagged English	Number sampled
Joomla	CMS (PHP)	278,821	172,989	1,000
TYPO3	CMS (PHP)	22,800	7,553	1,000
SilverStripe	CMS (PHP)	1,985	1,678	1,000
Drupal	CMS (PHP)	209,901	158,917	1,000
Xoops	CMS (PHP)	1,317	676	676
Plone	CMS (Python)	9,507	6,509	1,000
RoR	Web app framework (Ruby)	24,603	13,781	1,000
PHP	Web scripting language	54,121	32,114	1,000
Apache	Group of software projects	28,3646	165,988	1,000
Mozilla	Group of software projects	30,9958	157,094	1,000
OpenOffice	Office software	35,718	11,185	1,000
libreoffice	Office software	12,905	1,969	1,000
Sum		1,245,282	730,453	11,676

2 Analyzing OSS Microblogging

2.1 Selection of Keywords/Projects

Six web content management systems (CMSes) were selected, five of them written in PHP (Joomla, TYPO3, Silverstripe, Drupal, and Xoops) and one written in Python (Plone).[1] CMSes were selected because they are tightly focused on a particular product used by specialists (mostly website developers and system administrators) but nevertheless have large enough user and developer populations to generate sufficient traffic. Furthermore they can be compared to each other.

The server-side web scripting language PHP and the web application framework Ruby on Rails (RoR) were added in order to compare communications regarding web development using PHP and Ruby.

OpenOffice and libreoffice were included because, compared to the CMSes, they were likely to have more end-users who were not IT professionals. Their inclusion would thus help to highlight the characteristics of microblogging in the more specialist projects.

[1] Statuses were also gathered for the keywords 'Geeklog", 'Mambo", 'mojoPortal" and 'WebGUI" in order to analyze the PHP-based content management system of those names, but the keywords were abandoned due to the small proportion of relevant statuses (in the case of Mambo) and the small number of statuses retrieved (in the other cases).

Apache and Mozilla, two umbrella organizations for a number of open source projects, were included because, while being similarly Web-centered to the CMSes, their wider focus promised to show us differences between communication in individual projects and larger open source organizations.

2.2 Selection of Microblogging Service

While Twitter is the most well-known microblogging service, there is an alternative service, identi.ca, which uses open source software and, unlike Twitter, allows users to import and export their data using the FOAF standard. It might be the case that open source users and developers would make greater use of identi.ca for ideological reasons. In order to check whether this is the case, the search APIs of both services were queried for the selected keywords/projects between 15 February and 6 March 2013. The results are shown in Table 2. The results clearly show several orders of magnitude more activity regarding open source keywords on Twitter compared to identi.ca, which justifies the selection of Twitter as the data source for this study.

Table 2. Numbers of Statuses Retrieved from Twitter and identi.ca Search APIs, Feb 15-Mar 6, 2013

Project/keyword	Twitter	identi.ca
Joomla	34,716	10
TYPO3	2,422	0
SilverStripe	457	1
Drupal	24,591	65
Xoops	1,447	0
Plone	1,324	3
RoR	39,916	5
PHP	77,114	875
Apache	43,280	103
Mozilla	41,077	303
OpenOffice	4,160	21
libreoffice	6,567	357
Total	277,071	1,743

2.3 Data Collection

The Twitter Search API was queried approximately every two hours from 7 May to 26 December 2012 for keywords related to the 12 OSS projects. The keywords were not mutually exclusive, so for example a status containing both "PHP" and "Drupal" could be included in both samples.[2] The statuses were saved to a PostGreSQL database. A total of 1,245,282 statuses containing the 12 keywords were collected, as shown in Table 1.

[2] In fact, 10 statuses occurred in samples for two keywords/projects.

After collecting the statuses, a random sample of English-language statuses for each keyword was exported from the database for manual coding.[3] A sample size of 1,000 statuses for each project was coded where available, giving a total of 11,676. This compares with the total of 3,152 Twitter posts examined by Ehrlich and Shami, although their sample was not divided into 12 projects.

2.4 Data Cleaning

The sampled data was cleaned by excluding the following kinds of status:

Non-English statuses. The number of non-English statuses in the sample was very low because the sample included only statuses tagged as English language.

Irrelevant statuses. The number of irrelevant statuses was also generally very low, with the exception of Apache, Mozilla and RoR. "Apache" obviously has many other uses, while "RoR" refers not only to Ruby on Rails but also to such things as the Retraining of Racehorses. There were a large number of irrelevant statuses containing the word "Mozilla" because when Firefox users tweeted titles of web pages on any topic the text would include "Mozilla Firefox."

Robot statuses. Many studies of microblogging have focussed on user behavior and accordingly have selected users who were recognizably individual people. Considerable numbers of Tweets, however, are generated automatically. This is also the case where open source software is concerned, and it is a thorny question whether to include them or not. It was decided to exclude these "robot" statuses, above all because they were much more numerous for some keywords/projects than others, making it difficult to compare results. Defining and identifying automatically generated statuses is not easy, but the following categories of statuses were excluded:

1. Machine status Tweets e.g.
   Fedora [f18-arm] :: [97.44%] Completed – [0] Built – [3] Failed :: Task Error [perl-OpenOffice-UNO-0.07-3.fc17]-[1142021]
2. Repository-generated statuses e.g.
   cesag committed revision 1694 to the Xoops France Network SVN repository, changing 1 files: cesag committed revi... http://t.co/nXREVkxj
3. CMS change statuses generated by crawlers detecting changes in web page code e.g.
   http://t.co/YhTLHvCe: Change from NetObjects Fusion to TYPO3 http://t.co/YhTLHvCe #cms
4. Statuses sent automatically when someone posts to a forum e.g.
   XOOPS: Re: Linux Xoops white page issue? [by kidx] http://t.co/MxW4r2Ba
5. Statuses generated by job sites. Based on examination of the samples, these were defined as posts with 'job' or 'elance' (from 'freelance') in the sender's name.

Duplicate statuses. Duplicate statuses, which were defined as those containing identical text (with any urls excluded) and posted by the same user. Statuses containing identical text but sent by different users were labelled as retweets.

Table 3 shows details of data cleaning.

[3] 59% of the statuses collected were tagged as being in English.

Table 3. Details of Data Cleaning

Keyword/ Project	Number sampled	Irrelevant	Not English	Robots	Duplicates	Number after cleaning
Joomla	1,000	1	1	323	27	650
TYPO3	1,000	0	3	45	1	951
SilverStripe	1,000	0	3	121	23	854
Drupal	1,000	1	4	147	9	839
Xoops	676	9	39	164	55	417
Plone	1,000	20	7	196	5	772
RoR	1,000	218	7	78	24	673
PHP	1,000	6	8	297	20	674
Apache	1,000	565	5	148	6	277
Mozilla	1,000	236	34	5	0	725
OpenOffice	1,000	0	18	31	49	902
libreoffice	1,000	0	33	266	18	696
Sum	11,676	1,056	162	1,821	237	8,430

2.5 Data Coding

Ehrlich and Shami [6], building on work by Java et al.[7] and by Zhao and Rosson [8], proposed that microblog posts can be sorted into six categories. This study employed Ehrlich and Shami's scheme, which is introduced below along with examples from our data. User names have been changed.

1. Status (giving details of the poster's current activities):

   "Doing some module troubleshooting on the @Xoops_forums #XOOPS #imAwesome :P"

2. Provide information (sharing information/URLs, reporting news)

   "Get an overview and demonstration of #Acquia Search: http://t.co/9I352KgY #Drupal"

3. Directed posts (addressed to one or more other users):

   "@userA Ah, well on TYPO3 Sonar you can't - that's decided by the profile. If you want to change it you'd need your own Sonar install imo."

4. Retweets:

   "RT @userB: PC World calls out #Plone as one of 10 award-winng apps to try: http://t.co/WOK9mU4Y"

5. Ask question:

   "CodeKit users: can I selectively add files from a project? I really don't need it to monitor my entire SilverStripe install, just templates."

6. Directed with question.

> "@userC im just getting into Web Development, but i dont if it would be more begginer firiendly to learn PHP or RoR, help!?"

It was not always easy to decide which category a post should fall into; making a distinction between status and provide information proved particularly difficult, e.g.

> "In thinking caps on Mozilla writeable society session. A few mins, great minds and awesome ideas. See 'em here: http://t.co/dHk1zVSk #pdf12"

In such cases, we put the Tweet into the current status category, which we broadened to include notices and reports of events happening on the day of posting e.g.

> "Great RoR Meetup tonight at @manilla office with @userD and lots of swearing. Good times."

The coding was carried out by the author alone; clearly it would have been better to have three or more coders in order to reduce error and bias. Table 4 shows the numbers of statuses in each category, and Figure 1 displays the same data as proportions.

We also broke down the large "provide information" category into domain-specific subcategories. Table 5 and Figure 2 show the numbers and proportions of subcategories respectively.

3 Results and Analysis

3.1 Status Updates versus Information Provision

Do OSS-related statuses follow general Twitter usage in concentrating on the user's current activities, or do they tend to provide more general information?

Motivation In their study of general Twitter users [9], Naaman et al. used cluster analysis to suggest that 80% of the users sampled were "Meformers", who mostly tweeted about what they were doing, and only 20% were "Informers" forwarding information. In contrast, Ehrlich and Shami [6] found that about 10% of tweets sent by their regular users in IBM were about current status, compared with just under 30% that were providing information. We can therefore hypothesize that, given a more professional context, OSS-related tweets will have a higher proportion of information provision than "Me now" content. However, this is not to deny the potential value of "Me now" posts as a lubricant maintaining easy communication between participants in OSS projects.

Results Percentages of current status updates were between 0.9% and 8.5% (see Table 4 and Figure 1), while those providing information accounted for between 36% and 73% of each sample.

Table 4. Numbers of Statuses in Each Category

Keyword/ project	Ask question	Directed	Directed with question	Provide info	Retweet	Current status	Sum
Joomla	3	28	7	455	146	11	650
TYPO3	57	75	23	321	419	56	951
SilverStripe	9	52	9	381	371	32	854
Drupal	12	71	12	435	272	37	839
Xoops	2	5	2	240	163	5	417
Plone	11	42	8	277	400	34	772
RoR	13	92	13	417	109	29	673
PHP	6	28	3	455	176	6	674
Apache	8	12	4	151	93	9	277
Mozilla	3	30	5	439	224	24	725
OpenOffice	27	132	21	363	285	74	902
libreoffice	34	119	27	259	198	59	696
Sum	185	686	134	4,193	2,856	376	8,430

Table 5. Subcategories of "Provide Information" Statuses

Keyword/ project	Book	Code release	Event	Documen- tation	General	Jobs	News- letter	Security	Sum
Joomla	8	42	9	10	208	172	1	5	455
TYPO3	3	40	61	10	182	10	13	2	321
SilverStripe	3	86	39	18	175	18	0	42	381
Drupal	14	34	56	52	160	117	0	2	435
Xoops	1	24	0	101	85	15	4	10	240
Plone	18	37	78	14	106	9	0	15	277
RoR	0	3	9	10	67	328	0	0	417
PHP	11	10	3	21	87	312	0	11	455
Apache	6	22	5	22	72	12	0	12	151
Mozilla	0	28	9	6	384	0	0	12	439
OpenOffice	8	36	4	8	288	4	0	15	363
libreoffice	0	32	5	24	183	2	0	13	259
Sum	72	394	278	296	1,997	999	18	139	4,193

The two office applications both show a higher proportion of current status posts than the other projects. Reading the statuses did indeed suggest a lot of general users posting tweets along the lines of

> "man im tryna type this paper but for whatever reason my OpenOffice is gone!"

This supports the hypothesis that statuses written by professionals related to their work tend to be more information providing than "Me now"-ish.

A number of factors help to explain the higher proportions of information providing Tweets in our samples than those found by Ehrlich and Shami. First,

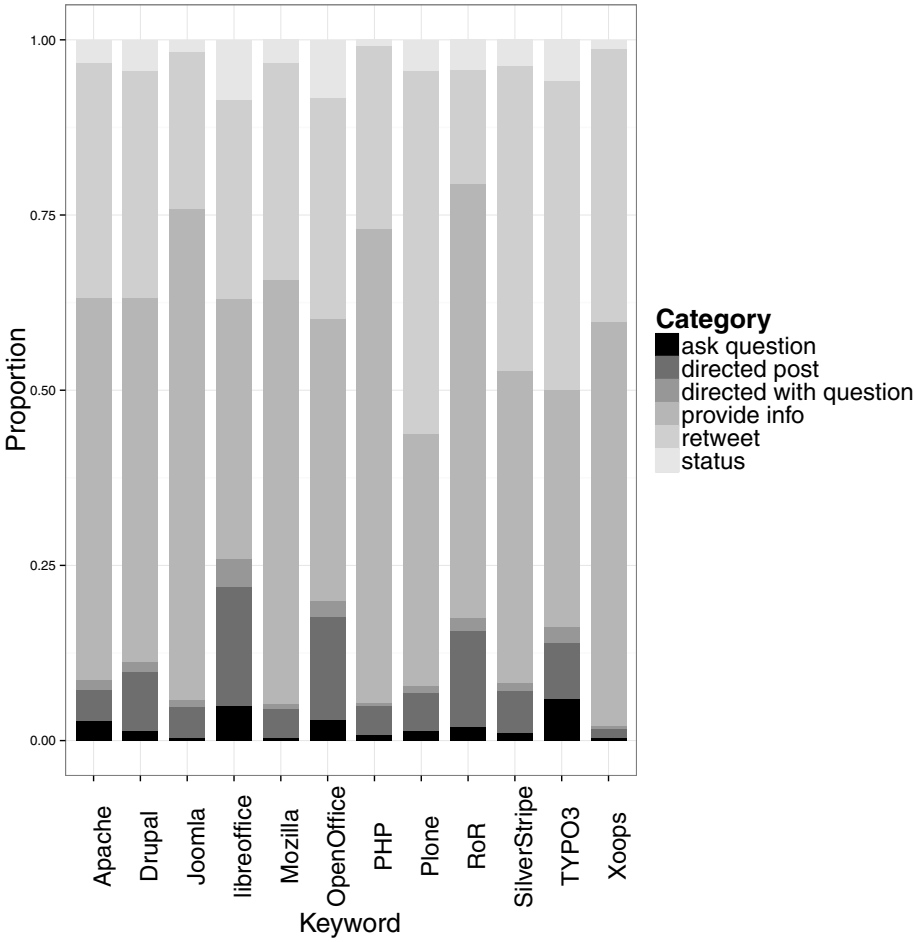

Fig. 1. Categories of Twitter Status

despite excluding many automatically generated job-related posts at the data cleaning stage, 1035 job-related posts remain in the "provide information" category; it is unlikely that Ehrlich and Shami's subjects would be sending many such posts. Second, the "provide information" category includes advertisements for online resources such as articles, books or software, which are also less likely to be sent by Ehrlich and Shami's subjects. Third, Naaman et al. found that tweets sent from mobile terminals are more likely than those sent from computers to be "Me now" rather than "provide information", and we can expect that most of the users and developers of the projects studied will be working on computers. Fourth, Naaman et al. also found that women rather than men are more likely to post "Me now" statuses, and given the high proportion of male participation in most open source software projects this may be having some effect.

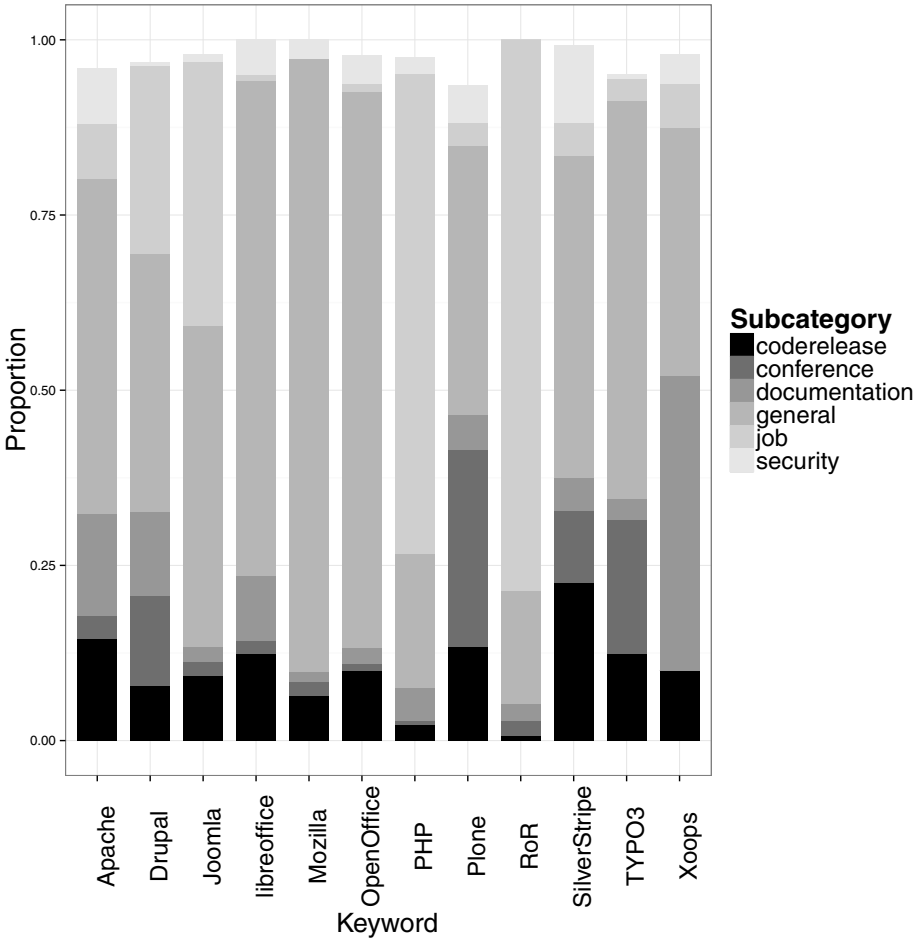

Fig. 2. Main subcategories of "provide information" statuses

3.2 Event-Driven Traffic

How much activity is prompted by events, both off- and online?

Motivation. Research on microblogging has shown that much activity occurs around events, particularly but not confined to offline events such as conferences [10, 11]. Vega et al. [12] found that conference participants tweeted more than usual in the week of the event. Conferences, and smaller events such as sprints and meetups, are the site of intense discussions between participants on all aspects of their project, so we can expect greater Twitter activity to exchange information with other participants and also to communicate what is happening.

Results. Table 5 and Figure 2 show the numbers and proportions of event-related tweets for the 12 project keywords.

Proportions of event-related tweets are low for all keywords except TYPO3 and Plone. In the case of TYPO3, there was something of a spike in activity around the conference held in Stuttgart in early October 2012, while for Plone there was a more pronounced concentration of tweets around the conference held in Arnhem in the same month. While it is not easy to distinguish current status and provide information categories concerning events, for neither project did tweets sharing current status e.g.

> "home and awake after a cool (exhausting) #T3CON12DE - now let's rock #TYPO3, #NEOS and everything like we rocked this conference!"

predominate, suggesting that communication is more focused on providing information about the conference to participants and non-participants e.g.

> "Thanks for attending our talk on #TYPO3 and #TYPO3Neos at #drupalhagen. Our slides are available here https://t.co/j6RJolxx."

It is interesting to note that for both TYPO3 and Plone, when we exclude job-related statuses, we also find that a higher proportion of statuses for these two projects mention "community" than is the case with other projects.[4] Only 4 out of the 35 and 7 out of the 56 statuses for Plone and TYPO3 respectively that mentioned community were event-related, e.g.

> "The #Plone community is just fucking crazy (in a good way). Prost!!!! #beersprint"

Therefore this does not seem to be merely a case of people tweeting about community when surrounded by their fellow developers and users. Further investigation is required to clarify whether this high proportion of event-related statuses and mentions of community is a coincidence, and if not then what is the relationship between the two. While it goes without saying that talking about community does not create one, the spontaneous nature of Twitter makes it a promising medium to explore how open source users and developers think and feel about their projects.

3.3 Rival Products

To what extent do statuses mention non-code related factors such as rival products?

Motivation. In their study of GNOME mailing lists, Shibab et al. [13] found that external factors, and particularly the emergence of rival products, played a significant role in shaping discussions among developers. They used these external developments to explain a decline in the market share of the Evolution

[4] We choose to ignore the large number of mentions of "community" for PHP because 24 of the 27 mentions are due to retweeting of a status praising one company's community engagement. The lack of comments on the retweets suggested an advertising campaign.

Table 6. Frequently occuring words in non-job related statuses

Project/ keyword	number of statuses	word	number of occurrences
Joomla	561	wordpress	106
		drupal	31
TYPO3	979	(none)	
Silverstripe	921	wordpress	38
Drupal	765	wordpress	75
		joomla	44
Xoops	525	(none)	
Plone	950	wordpress	36
RoR	368	php	23
PHP	374	apache	110
		mysql	83
		ruby	30
		rails	30
		magento	23
Apache	406	mysql	137
		openoffice	30
Mozilla	729	google	75
		chrome	45
		windows	41
		ipad	32
		microsoft	28
OpenOffice	923	libreoffice	94
		excel	93
		word	68
		microsoft	42
libreoffice	946	openoffice	68
		office	68

mail client as rival products emerged. We can expect that a similar analysis of Twitter activity will show which products are perceived as rivals by those within and without open source projects. These findings, particularly if they could be tracked over years rather than months, would help to explain design decisions and shifts in market share.

Results. We excluded job-related statuses, then counted the occurrences of words in the text of statuses for each keyword/project. Table 6 shows a selection of the technology-related words appearing in the top 50 most commonly used words for each project, along with the number of statuses analyzed. Note that the number of occurrences can be greater than the number of statuses because project names are often used more than once in a single tweet. Some of the words are components or closely related to the projects and some are rival products.

Of the CMSes, statuses regarding all except TYPO3 and Xoops mention rival products, predominantly Wordpress. It would be worthwhile to follow this up with a longitudinal study to establish whether Wordpress is gaining in its

position as the chief rival to most open source CMSes. Tweets about the office products, as expected, mention each other and Microsoft Office. It is also interesting to see that statuses on PhP and Ruby on Rails both make significant mention of each other.

4 Threats to Validity and Topics for Further Research

One limitation of this study is that it compares proportions of different kinds of Twitter use across projects/keywords based on samples of similar size, while the absolute numbers of statuses for the different keywords differ greatly. Therefore, while we can conclude that e.g. a higher proportion of Silverstripe-related tweets than Apache-related tweets are concerned with events, that does not mean that Apache users and developers do not make equally active use of Twitter with regard to events, while also using Twitter more actively to e.g. provide links to documentation.

It would be desirable to increase the sample size for each project in order to increase the reliability of the data. It would also be better if the sample came from a full twelve-month period because many projects have large conferences once a year, and the current seven-month period risks excluding some events.

Not all statuses containing those keywords during the period were collected, for two reasons. First, the Twitter Search API does not guarantee to return all statuses for a given period. Second, it was not possible to ensure that the data collection script ran uninterrupted for the entire period. When such interruptions occurred, and the volume of posts was high, due to the limitation placed on the number of results returned from Twitters Search API (1500 in this case), it was not possible to collect all the statuses posted since the previous collection. These gaps might introduce distortions into the findings, for example if they coincide with a major conference or code release related to a particular project, thus missing a spike in microblogging activity. However, the long period of data collection can be expected to reduce the impact and perhaps to even out such distortions. The incompleteness of the data would also be a problem if the study were aiming to do a network analysis of Twitter-based communication in OSS projects, as there would be many missing directed messages and replies. However, as our purpose here is merely to analyze the content of individual statuses, this is not an issue for the current research.[5]

In retrospect, it would have been desirable to include Wordpress, one of the most popular open source CMSes. Unfortunately Twitter's search API does not allow statuses more than a few days old to be collected.[6]

The study could also be improved by obtaining information about numbers of followers and giving greater weight to posts that are more read.

[5] In addition, gathering only statuses that contain particular keywords would not capture all the directed Tweets sent between any given set of users.

[6] Some commercial services offer to retrieve historical Tweets.

References

1. Boehringer, M., Richter, A.: Adopting enterprise 2.0: A case study on microblogging. Mensch & Computer 2009: Grenzenlos frei (2009)
2. Reinhardt, W.: Communication is the key-support durable knowledge sharing in software engineering by microblogging. In: Proc. of the SENSE Workshop, Software Engineering within Social Software Environments (2009)
3. Riemer, K., Richter, A., Bohringer, M.: Enterprise microblogging. Business & Information Systems Engineering 2(6), 391–394 (2010)
4. Guenther, O., Krasnova, H., Riehle, D., Schoendienst, V.: Modeling microblogging adoption in the enterprise. In: Americas Conference on Information Systems (AMCIS) 2009 Proceedings, vol. 544 (2009)
5. Riemer, K., Richter, A.: Tweet inside: Microblogging in a corporate context. In: Proceedings of the 23rd Bled eConference, pp. 1–17 (2010)
6. Ehrlich, K., Shami, N.: Microblogging inside and outside the workplace. In: ICWSM 2010 (2010)
7. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 56–65 (2007)
8. Zhao, D., Rosson, M.B.: How and why people twitter: the role that micro-blogging plays in informal communication at work. In: Proceedings of the ACM 2009 International Conference on Supporting Group Work, pp. 243–252 (2009)
9. Naaman, M., Boase, J., Lai, C.: Is it really about me?: message content in social awareness streams. In: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, pp. 189–192 (2010)
10. Vega, E.: Communities of Tweeple: How Communities Engage with Microblogging When Co-located. PhD thesis, Virginia Polytechnic Institute and State University (2011)
11. Ebner, M., Mühlburger, H., Schaffert, S., Schiefner, M., Reinhardt, W., Wheeler, S.: Getting granular on twitter: Tweets from a conference and their limited usefulness for non-participants. In: Reynolds, N., Turcsányi-Szabó, M. (eds.) KCKS 2010. IFIP AICT, vol. 324, pp. 102–113. Springer, Heidelberg (2010)
12. Vega, E., Parthasarathy, R., Torres, J.: Where are my tweeps?: Twitter usage at conferences. Paper, Personal Information Management class, Virginia Polytechnic Institute and State University (June 1) (2010), `http://www.socialcouch.com/demos/final_paper_twitter.pdf`
13. Shibab, E., Bettenburg, N., Adams, B., Hassan, A.E.: On the central role of mailing lists in open source projects: An exploratory study. In: Proceedings of the 3rd International Workshop on Knowledge Collaboration in Software Development, KCSD, Kyoto, Japan (November 2009)