Les participants à Stack Overflow travaillent-ils sur le long terme ?

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 14 juin 2009

https://www.bortzmeyer.org/stack-overflow-short-tail.html

Le site social "*Stack Overflow*" <https://www.bortzmeyer.org/stack-overflow.html> ayant désormais rendu publique sa base de données <https://www.bortzmeyer.org/stackoverflow-to-postgresql.html>, il est possible de l'analyser et d'étudier les comportements des participants à ce site. Par exemple, les votes et les réponses se poursuivent-ils pendant des mois ou bien sont-ils concentrés dans les heures qui suivent la publication d'une question ?

Regardons d'abord l'évolution des votes dans le temps. Combien de temps s'écoule t-il entre un message (question ou réponse) et les votes sur ce message ?

```
so=> SELECT DISTINCT (votes.creation - posts.creation::date) AS interval,
so->                 to_char(count(votes.id)*100.0/
so(>                         (SELECT count(votes.id) FROM Posts, Votes
so(>                                 WHERE Votes.post = Posts.id), '99.99') AS
so->                  percent
so->    FROM Votes, Posts
so->    WHERE Votes.post = Posts.id
so->    GROUP BY interval ORDER BY interval;
 interval | percent
----------+---------
        0 |  54.75
        1 |  10.33
        2 |   2.62
        3 |   1.74
        4 |   1.26
        5 |   1.01
        6 |    .87
        7 |    .70
        8 |    .57
        9 |    .49
       10 |    .44
       11 |    .42
       12 |    .38
       13 |    .39
       14 |    .39
       15 |    .33
       16 |    .30
...
```

La majorité absolue des votes ont lieu le jour de publication de l'article. Dès le cinquième jour, on passe en dessous d'un pour cent des votes pour la journée. Bref, il semble que le syndrôme connu à "*Stack Overflow*" sous le nom de "*fastest gun in West*" <http://stackoverflow.uservoice.com/pages/general/suggestions/24695> joue à plein. N'importe quel contributeur peut le voir : si on arrête d'écrire, la réputation ne bouge plus guère, les votes tardifs étant rares.

Cela se voit très bien sur le graphique, produit par gnuplot avec ces instructions :

```
set xrange [:150]
set format y "%.0f"
plot "votes-per-day.dat" using 1:2 with lines title ""
```

et en extrayant les données avec `psql --file votes-per-day.sql --field-separator ' ' --tuples-only --no-align so > votes-per-day.dat`, on obtient :

Mais la traîne est tellement longue qu'elle contribue quand même. Ainsi, 27 % des votes ont lieu plus d'une semaine après l'article et 19 % des votes plus d'un mois après :

```
so=> SELECT count(votes.id)*100/(SELECT count(votes.id) FROM Posts, Votes
                                 WHERE Votes.post = Posts.id) AS percent
       FROM votes, posts WHERE
                   votes.post = posts.id AND
                   votes.creation >= (posts.creation + interval '7 day')::date;
 percent
---------
      27
(1 row)

so=> SELECT count(votes.id)*100/(SELECT count(votes.id) FROM Posts, Votes
                                 WHERE Votes.post = Posts.id) AS percent
       FROM votes, posts WHERE
                   votes.post = posts.id AND
                   votes.creation >= (posts.creation + interval '30 day')::date;
 percent
---------
      19
(1 row)
```

Graphiquement, en mettant l'axe des Y en logarithmique, on voit mieux la longue traîne. Les instructions gnuplot sont :

```
set xrange [:150]
set logscale y
set format y "%.0f"
plot "votes-per-day.dat" using 1:2 with lines title ""
```

et le résultat est :

Et l'intervalle entre la question et une réponse **acceptée** (dans "*Stack Overflow*", l'auteur d'une question peut marquer une question et une seule comme acceptée) ?

```
so=> SELECT DISTINCT(answers.creation::date - questions.creation::date) AS interval,
so->        to_char((count(questions.id)*100.0/(SELECT count(answers.id)
so(>                                  FROM Posts questions, Posts answers
so(>                                  WHERE answers.id = questions.accepted_answer AND
so(>                                        questions.type = 1 AND
so(>                                        answers.type = 2)), '999.9') AS percent
so->     FROM Posts questions, Posts answers
so->     WHERE answers.id = questions.accepted_answer AND questions.type = 1 AND
so->           answers.type = 2
so->     GROUP BY interval ORDER by interval;
 interval | percent
----------+---------
        0 |   83.9
        1 |    7.3
        2 |    1.5
        3 |    1.0
        4 |     .7
        5 |     .5
        6 |     .4
        7 |     .4
        8 |     .3
...
```

La courbe est encore plus brutale. On pourrait résumer en disant que, si on ne répond pas le premier jour, on n'a que peu de chances d'être accepté... Mais il faut garder espoir. 4 % des réponses acceptées ont été écrites plus d'une semaine après l'article et 1 % après un mois.

```
so=> SELECT count(answers.id)*100/(SELECT count(answers.id)
                                FROM Posts questions, Posts answers
                                WHERE  answers.id = questions.accepted_answer AND
                                       questions.type = 1 AND
                                       answers.type = 2) AS percent
       FROM Posts questions, Posts answers
       WHERE answers.id = questions.accepted_answer AND
             questions.type = 1 AND answers.type = 2 AND
             answers.creation > (questions.creation + interval '7 day')::date;
 percent
---------
       4
(1 row)

so=> SELECT count(answers.id)*100/(SELECT count(answers.id)
                                FROM Posts questions, Posts answers
                                WHERE  answers.id = questions.accepted_answer AND
                                       questions.type = 1 AND
                                       answers.type = 2) AS percent
       FROM Posts questions, Posts answers
       WHERE answers.id = questions.accepted_answer AND
             questions.type = 1 AND answers.type = 2 AND
             answers.creation > (questions.creation + interval '30 day')::date;
 percent
---------
       1
(1 row)
```