Applying a User-Centered Metric to Identify Active Blogs

Adam D. I. Kramer
Department of Psychology, University of Oregon
adik@uoregon.edu
http://www.uoregon.edu/~adik/

Kerry Rodden
User Experience Research, Google
kerryr@google.com
http://www.rodden.org/kerry/

Introduction: Blog Abandonment

• Blogging is popular: BuzzMetrics estimates over 38m blogs.
• However, many people stop posting to, or "abandon" their blog.
• Separating the active from the abandoned matters:
  - Why do people abandon blogs? How can we retain bloggers?
  - Do abandoned blogs retain readers? Which ones do or don't?

## Research Question: How can we detect blog abandonment?

• Time since last post, e.g. "no posts within the last 30 days"
• Posts since a certain epoch, e.g. "no posts since October 2005"
  - But what of "monthly bloggers?" What of "fad" bloggers?
• Qualitative methods (humans reading and coding blogs)
  - But variable in terms of correctness; time-consuming

A User-Centered Activity Metric

• Different people have different post habits
• But each person's posting habits should be reasonably coherent
• Can we build an "activity" metric based on individual habits?

Method

• A small subset (n=1,100,810) of Blogger blogs were analyzed
• Individual posting habits were tracked:
  - Total number of posts (through June 28, 2006)
  - Lifespan (number of days between first and last post)
  - Time between posts (incl. mean and variance)
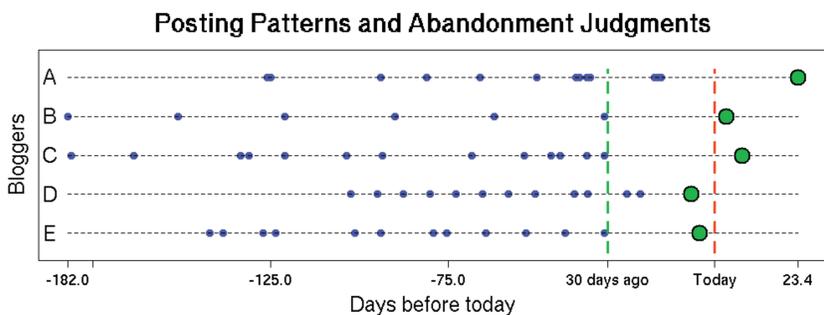
Step 1: Identify "Established" Blogs

• Many blogs showed very few posts, or a very short lifespan
  - These blogs never "took off;" never became "established."
• We removed blogs with less than a 9-day lifespan
  - Enforces blogging through at least one week and weekend
• We removed blogs with less than 11 total posts
 - Allows us a robust estimate of the post rate variance

Results

• Established blogs have highly skewed mean times between posts
• Any metric based on time between posts (e.g., 30 days since last post) means something different for each blog

Histogram of time between posts, across bloggers
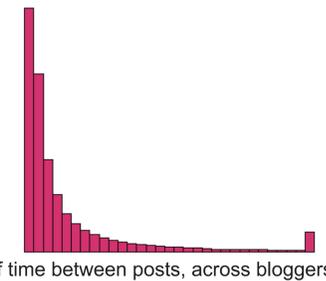
Step 2: Identify "Active" Blogs

• For established blogs, the times between posts were normally distributed **for each** blogger
  - This means that each blogger's behavior is coherent
  - But **not** normally distributed **across** bloggers!
• Thus, posting patterns are only coherent within persons
• **Conclusion: Activity can and must be estimated using idiographic (user-centered) parameters!**
• Given a normal distribution of time between posts, 99.4% of posts will be less than 3 standard deviations above the mean
 - We call a blog "abandoned" if the last post was more than *three standard deviations beyond that blogger's average delay.*

Test case: 30-day vs. Idiographic

• We compare our metric to the commonly-used 30-day metric
  - How many 30-day active blogs do we think are *unestablished* or *abandoned?*
  - How many 30-day active blogs do we think are still *active*?
  - How often do our metric and the 30-day active metric *agree*?

30-day Active Blogs

• 62% of blogs fitting the 30-day active metric were *rejected* for not being "established;" likely "fad" or spam blogs
• 6% of 30-day active blogs (see Blogger D below) are *abandoned* by our metric.
  - Though posted to this month, the delay in posting was uncharacteristic *of that blogger!*
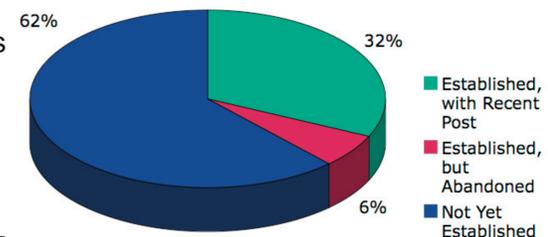
62% 32% 6%
Established, with Recent Post
Established, but Abandoned
Not Yet Established

30-day Inactive Blogs

• We *reclaim* 2% of blogs not meeting the 30-day activity metric (Blogger C, below).
  - For these bloggers, a gap of over 30 days is *predictable* and *characteristic* of their behavior!
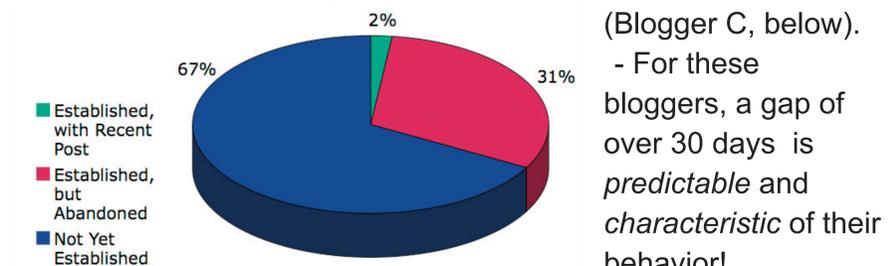
2% 67% 31%
Established, with Recent Post
Established, but Abandoned
Not Yet Established

- As most blogs are 30-day inactive, 2% is actually very large!

Summary

• Blog abandonment must be inferred from prior blog activity
• However, this metric agrees with popular metrics for most blogs
• For blogs with 11 or more posts, time between posts is normally distributed within bloggers (but not across bloggers)
• Thus, researchers can choose their own error rates to accompany the needs of various research projects
• Blog analysts should consider idiographic methods

Posting Patterns and Abandonment Judgments

•: Posts, •: Idiographic cutoff, - - -: 30-day cutoff, - - -: Today
• Predicted point of inactivity based on a user's individual behaviors is shown by the larger green dot, in comparison to the 30-day metric which is shown by the red line.

The Idiographic Metric Adapts to User Behavior

•: Posts; O,O: Idiographic "guesses", □,□: 30-day "guesses"
O,□: Correct guesses; O,□: Incorrect "guesses
• Our metric "learns" the user's habits over time
• Once a blog has at least 11 posts, our metric becomes stable