your.flowingdata: Personal Data Collection via Twitter

Nathan Yau

Department of Statistics

Center for Embedded Networked Sensing

University of California, Los Angeles

nyau@stat.ucla.edu

Mark Hansen

Department of Statistics

Center for Embedded Networked Sensing

University of California, Los Angeles

cocteau@stat.ucla.edu

Abstract

In this proposal we describe your.flowingdata, a system that allows users to collect personal data via Twitter and the more general implications it has for personal informatics.

Keywords

Personal informatics, Twitter, SMS, visualization, data collection

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Workshop proposal, personal informatics, Twitter, SMS, visualization, data collection

Introduction

People use Twitter to update others on what they are doing with messages, or *tweets*, that are 140 characters or less. Upates range from the eventful, like the birth of a new child, to the mundane, like what someone just ate. In any case, a culture has developed around Twitter to send updates on what is going on at any given moment. your.flowingdata (YFD) [1] aims to

use this culture to intertwine personal data collection into peoples' everyday.

YFD lets people send private tweets, or *direct messages*, to collect personal data to both monitor short-term behavior and to update a data journal. People can then explore their data visually, look for patterns, try to make conclusions about their behaviors, and selectively share their findings.

Ultimately, as a personal data collection test bed that is constantly in-development, YFD provides insights on what such a system might need to be useful to the individual.

Data Collection

Users collect data by sending direct messages in a specified format to @yfd, a Twitter account whose only function is to receive these messages. Because this is via Twitter, users can send updates via a variety of desktop clients, browser, or SMS. Again, the goal of YFD is to intertwine personal data collection with regular usage of Twitter, so this plays to our advantage. The data-logging syntax was also designed with this in mind. Here is the basic format:

```
<action> <value (optional)> <unit (optional)>
```

The `<action>` is usually a verb such as *consumed* or *watched*. The `<value>` is numeric and the `<unit>` is a string representing a category or unit of measurement. For example:

```
ate 2 hot dogs
```

The `<action>` is *ate*, the `<value>` is *2* and the `<unit>` is *hot dogs*.

Further extensions of the syntax let users timestamp their data by appending time to the end of their direct message as well as tag their data with hashtags, another Twitter convention.

Visualization and Exploration

Another challenge of YFD is to provide an interface that lets users explore their data in a way that is familiar but lets them dig deeper than static charts. YFD provides a growing set of interactive visualization tools. Some visualization will look famliar such as the interactive treemap or the Action Explorer, made popular by Martin Wattenberg's work with baby names [2].
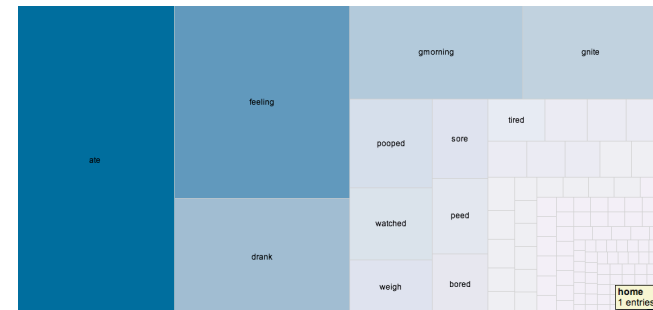

figure 1 An interactive and searchable treemap allows the user to explore proportions.
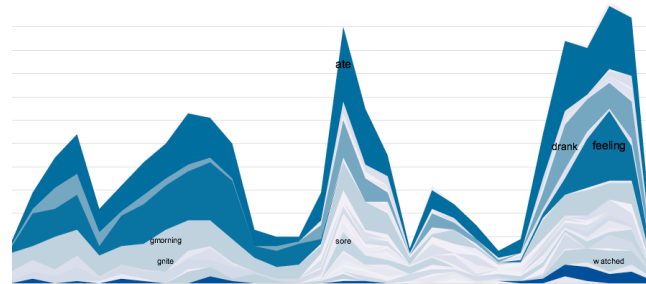
figure 2 A stacked area graph allows exploration of trends.

There are also more experimental visualizations that were built specifically for the data at hand. First there is a tool to look at cross-correlation. While rough, the visualization helps users find actions that are related in some way. Users can also explore durations between a start and stop actions such as *gmorning* and *gnight*.
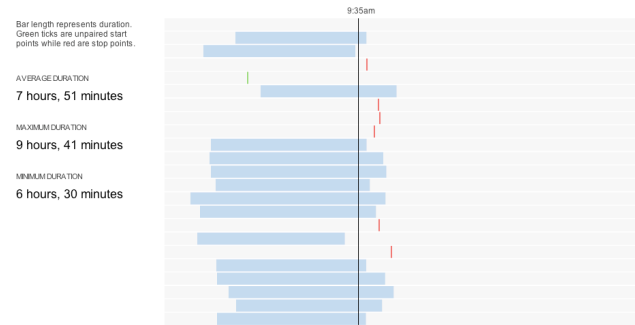

Bar length represents duration. Green ticks are unpaired start points while red are stop points.

AVERAGE DURATION
7 hours, 51 minutes

MAXIMUM DURATION
9 hours, 41 minutes

MINIMUM DURATION
6 hours, 30 minutes

9:35am

figure 3 The user can explore durations between two actions.

In an effort to provide familiarity, YFD also provides a calendar view, which displays color intensity based on the frequency of an action on any given day.
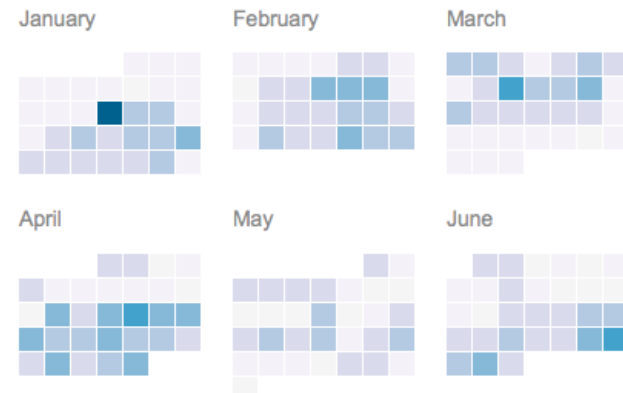

January February March

April May June

figure 4 A calendar view provides the user with a familiar interface.

Usage

While still very much in development, YFD has a healthy user base to test new features and perhaps provide insights on what a personal data collection system might need to be productive to the individual. At the time of writing this, 2,500 users have logged 194,000 data points. The bulk of data has been logged by about 650 users, who have logged more than twenty data points each, while other users did not go too deeply into the application. The YFD site gets about 45,000 pageviews per month.

Many users log typical data about themselves like what they eat, how much they weigh, and when they go to sleep. However, some have found others uses for YFD. Some collect data not about themselves, but dependents such as children and pets. One user collects weather data in his hometown. Another has

synchronized his music-listening habits on last.fm. Some use Twitter's API to automate the collection process. Finally, one user developed an iPhone application to make data entry easier. It is freely available in the iTunes store.

With all of that said, and given the design of YFD, one might think that the most active YFD users are those who are active Twitter users; however the separation is not so clear-cut. Some active YFD users update their Twitter feeds rarely, while others update Twitter feeds way more than YFD. The same goes for non-active YFD users. For example, the median number of Twitter updates for YFD users who have logged less than 20 actions is 535 while the median for those who have logged more than 100 actions to YFD is 387.

Discussion
Clearly, there is still a lot of work to be done. YFD has experimented with reminders, alerts, and a query system, but current usage suggests there seems to be a need for automation in both the data collection step and analysis step.

There is also a pedagogical consideration. How do we teach non-professionals how to analyze data? The Personal Environmental Impact Report (PEIR)[3] for example, was a previous project that focuses on a different dataset but aims to repurpose mobile phones to estimate an individual's carbon impact on her surroundings. Traditional statistical visualization such as histograms was often unfamiliar to users; however, a map that showed GPS traces proved easier to read.

A statistician knows or can figure out where to look in a dataset to find trends, outliers, and patterns, but how does a non-professional come to similar conclusions in an informal setting? How can we teach users to ask the right questions about their data? Can non-professionals understand concepts beyond count, average, and the obvious trends? Hopefully, future YFD development can provide some answers.

Citations
[1]   your.flowingdata, http://your.flowingdata.com/.

[2]   Baby Name Wizard. http://babynamewizard.com/.

[3]   M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, P. Boda. PEIR, the Personal Environmental Impact Report, as a Platform for Participatory Sensing Systems Research, MobiSys, June 2009.