Large Scale Analysis of Web Revisitation Patterns

Eytan Adar, Jaime Teevan, Susan Dumais
University of Washington
Microsoft Research

50-80% of Web page visits

(Should) impact:
- Browser design
- Search engines algorithms
- Site design

What makes one page’s revisits different from another?
Prior work

- Cockburn, Greenberg, McKenzie, Tauscher, (IJHCS’97, JIS’00, IJHCS’01)
- Kellar et al. (CHI’07, ASIST’06)
- Obendorf et al. (CHI ’07)
- (many others surveyed in paper)
Data
- Dual study
- Usage Logs
  - Live Toolbar (August 2006, 5 weeks)
  - Around 600k users
    - Subset of total
- Sampled URLs
  - Around 55k
- User study

Data
- Dual study
- Usage Logs
- User study
  - 20 users
  - Access to any of the sampled URLs
    - Custom built logging tool
  - Survey
    - Why did you visit?
    - How often do you think you visit? etc.

Sampling URLs
- 523 unique users
- Unique Users (popularity)

Sampling URLs
- Visits Per User
  - User 1 = 10 times
  - User 2 = 12 times
  - ...
- Unique Users (popularity)
Normalization

If you learned nothing else...

Lots of people revisit the website at this time (the peak)

1 hour

More than average revisit every 1 hour

If you learned nothing else...

Types of Revisitiation Curves

- Clustering algorithm on the 55k URLs
  - 12 clusters, 4 main groups

Fast

Medium

Slow

Hybrid
4 Main Types

- Fast < Hourly
- Medium Hourly, Daily
- Slow >Daily/Weekly

4 Main Types

- Fast < Hourly
- Medium Hourly, Daily
- Slow >Daily/Weekly
- Hybrid < Hourly & > Daily/Weekly
Session
- Which page was looked at before page of interest
  - Same page?
  - Same domain?
  - Search?

Content
- Open directory labels (ODP/DMOZ)
  - Arts, Games, Business, etc.
- Safety labels (porn, kids, etc)
- Specific terms in text + URLs

Session
- Fast
  - Hub and spoke
    - Shopping menu (1 time)
  - Auto refresh
    - 77% from same domain
      (versus 43% for medium)
- Medium
  - Also search (4.0%)
- Slow
  - Search before access (4.3% versus 2.9%)
- Hybrid
  - Shopping menu (regular)
  - "Catch up"
Content

- Fast
  - Porn & spam, shopping, reference
- Medium
  - Mail, “login,” games
- Slow
  - Specialty search, kids, weekend entertainment
- Hybrid
  - Entertainment & hobbies, Craigslist, “login”

Intent

- User Survey
- Fast
  - Buy something, monitor live content (traffic, sports, etc.)
- Medium
  - Communicate, listen to music, view videos, play games
- Slow
  - Interact with personal data, view previously viewed information (refinding)
- Hybrid
  - Visit new content or follow new links, buy something

Rate of Change

- Crawled once an hour for a month
- Fast
- Medium
- Slow
- Hybrid

Average page changes in a month

- Further work: finer granularity, nature of change

Browser Implications

- Monitoring Data changes and users want to keep up
- Further work: finer granularity, nature of change
Browser Implications

Fast
- Working stack or bread crumbs
- Fast vanishing

Medium
- Regular access
- Notify user?

Slow
- Infrequent access
- Hard to find?

Search Engines

- Crawlers
  - Speed related to popularity and change
  - Should adapt to revisits
    - Indication of interest
- Improved ranking
  - Based on revisitation patterns
  - Clustering, etc.

Summary

- Very large scale study of revisitation
  - 600k users, 55k pages
- User study
- Revisitation curves
  - 4 main types
- Properties of revisits
  - Crawls, labels, etc.
- Implications to browsers, search (and more)

Future/Ongoing Work

- Detailed analysis of content
  - Change and resonance
  - Finer granularity
- Predicting revisitation behavior
- Prototype development
Questions?
Average Revisititation Curve

1 day
2 days
noisy

Average Curve

16 “Exponential” Bins

Binned data

fast

slow