Notes on blogspam and web search

Tim Converse
Yahoo! Web Search
Relevance Ranking Group

About me

- Who I am
  - I lead the algorithmic anti-spam group at Yahoo Web Search
  - Helped Brian and Marc organize AIRWEB 06
- Who I’m not
  - A blog search person
  - Andrew Tomkins

Outline - web search impact

- Comment spam and nofollow
  - Publicly-writable pages violate “good doesn’t link to bad”
- Fake blogs / splogs
  - Just a lower barrier to publishing entry
- RSS/Syndication ecology
  - Infrastructural support for scrape/copy/”aggregate”

Notes on nofollow

- Standard introduced about 1 year ago
  - Misnomer - untrusted?
  - Unprecedented websearch industry cooperation!
- My personal predictions at the time:
  - It won’t be widely adopted; It won’t help even if it is
- I was at least 50% wrong

Nofollow adoption

- About 1 in 100 links we see is nofollow-labeled
- Without revealing exact number: billions of nofollow links
- Personal opinion of net effect on blogosphere/web:
  - Four-door car with two doors locked
  - Successful attacks: discriminate, indiscriminate
  - Net effect probably positive

Splog observations

- Started later than comment spam, now more important
- Trend is toward splogs for link creation, rather than destination
- Marked variation by hosting service depending on 1) barriers to entry, 2) policing
- When is a blog hosting service a spam super-domain?

RSS and syndication

- Framework for scraping and (worse) “aggregating” fresh content for spam. Can’t determine original author.
- Spectrum between intelligent aggregation and spam scraping/weaving/stitching
- Research challenge: Spam page with two paragraphs, stolen from two different sources
- Prediction: increased auth between publisher and engine (Google SiteMaps, Yahoo! SiteExplorer)