What are splogs?

- Splog (Spam + Blog): a blog created with the intention of spamming
- Splogs are polluting the blogosphere
  - 10-20% of blogs are splogs
  - 75% of new pings come from splogs
  - 44 of the top 100 results from 3 top blog search engines are splogs
Example (1): keyword stuffing

- Specific keywords appear repetitively in blog title, post title, description, etc.
- Commercial intention: host Google AdSense
- Lots of out-going links pointing to itself

Example (2): stolen content

Computer science thesis

Sat, 17 Jun 2006 02:32:20 -0500

Lenovo & Dell--Support Makes the Difference
by Computer science thesis (computer-science-thesis) @ Thu, 15 Jun 2006 02:02:15 -0500
In world of seemingly undifferentiated products and services, support is everything. When a computer has a hiccup, and a user is on a hotline, they are scared and apprehensive. Their lives as they live them have stopped.

...}

Original post: Lenovo & Dell--Support Makes the Difference by at Google Blog Search: computer services

Thu, 15 Jun 2006 02:02:15 -0500

Man-waiChang
by Computer science thesis (computer-science-thesis) @ Thu, 15 Jun 2006 02:02:15 -0500

Trained in software engineering, information systems, computer hardware systems and formal specification. Attempted part-time MSc in Computer Science in City University of Hong Kong, but failed to complete the thesis owing to heavy ...

Data Recovery Free To Air TV Acne News Home Security News Computer News

Original post: Man-waiChang by at Google Blog Search: computer science thesis

Wed, 14 Jun 2006 02:03:53 -0500

C.2.1 Simulation Reporting Document Tutorial
by Computer science thesis (computer-science-thesis) @ Wed, 14 Jun 2006 02:03:53 -0500

There is a data reporting page for every year. It will look something like this: There is a lot of data on this screen, but it is organized logically. The following pages explain the details.

Computer Science Thesis Index.

Original post: C.2.1 Simulation Reporting Document Tutorial by at Google Blog Search: computer science thesis

Traditional content analysis is not enough
Example (3): link farm

- Post content is aggregated from other sources, may or may not relevant.
- Blogroll and Links list blogs in almost the same structure, all about sports and have affiliate links to one another, with fake (machine-generated) content.

Example (4): trackbackback links

- Traditional link analysis is not enough
What are splogs?

- Motive of creating splogs
  - drive visitors to affiliated sites that have profit-making mechanisms, e.g. Google AdSense or PPC
- Working definition
  - *any deliberate action that is meant to trigger an unjustifiably favorable relevance or importance, considering the blog’s true value*
- NOTE
  - evaluated at the blog level, not at individual pages
  - a blog having comment spam or trackback spam is *not* considered a splog

Problem Statement

**Goal**: combat spam in the blogosphere
- What are *splogs*?
- How to detect splogs?

**Approach**: using temporal regularity of *content, post time* and *links*
- Self-similarity analysis
- Visual characterization
- Temporal feature computation

Splog detection—different from web spam detection!
Related work

- **Web spam detection**
  - Link analysis
    - [Gyongyi05]: spam mass estimation
    - [Wu&Davison05]: determine spam seeds based on overlap between incoming and outgoing links
    - [Fogaras&Racz05]: multiple-step neighborhood searching
  - Content analysis
    - [Fetterly04][Ntoulas06]: statistical properties in content
    - [Fetterly05]: identifying phrase level replication
- **Splog detection**
  - [Kolari06]: apply web spam detection & topic identification techniques in splog detection
  - [Salvetti&Nicolov]: URL tokenization

**Approach**

1. Self-similarity analysis
2. Visual characterization
3. Temporal feature computation
Temporal structures in splogs

1. Self-Similarity Analysis

- Graph model
  - Nodes: posts from the same blogs
  - Edges: similarity measure
- Self-similarity matrix: generalized autocorrelation over any time series of media objects
  - Post time
  - Post content
  - Post links
Post time

\[ S_{\text{micro}}(i, j) = \min(|t_i - t_j|, \delta_{\text{micro}}) \]

\[ S_{\text{macro}}(i, j) = |t_i - t_j| \]

Daily time

Absolute time

Similarity matrix of post time

Post content / links

Histogram intersection

Identical contents

\[ S_c(i, j) = \frac{\sum_{k=1}^{M} \min(h_i(k), h_j(k))}{\sum_{k=1}^{M} \max(h_i(k), h_j(k))} \]

\[ S_l(i, j) = \frac{\sum_{k=1}^{M} \min(h_i(k), h_j(k))}{\sum_{k=1}^{M} \max(h_i(k), h_j(k))} \]

Retirement - California retirement community

Identical contents

Super Decoration Guide - Zen Decoration

Link to the same website

Similarity of time

Topic change

Content similarity

Link similarity
2. Visual Characterization

- Clock-like representation
  Each post is represented as a post-arrow
  - Orientation: post time
  - Length: similarity w.r.t. previous post

Different blog types
Examples of “splog clock”

3. Temporal Features

- Features along off-diagonals
  - High intensity off-diagonals: attribute values are similar at different posts

- Features from coherent blocks
  - High intensity blocks: attribute values remain highly constant for some period
**Joint features**

- From two attributes

**Experimental Results**

- Dataset
- Annotation tool
- Classification
- Detection Performance
Dataset

- TREC Blog Track 2006 Collection: 100,649 feeds
  - 100,649 feeds, collected over 11 weeks
  - Removing duplicate feeds, feeds without homepage or permalinks → 43.6K unique blogs
- Annotation interface
  - (N) Normal
  - (S) Splog
  - (B) Borderline
  - (U) Undecided
  - (F) Foreign
- Ground truth
  - Label 9167 blogs
  - 7380 normal, 897 splogs

Classification

- Binary classification: splog or normal blogs
  - Use SVMs classifier with a radial basis function kernel
  - Combine baseline features with new features
  - Baseline: A subset of content features in [Ntoulas06]
    - Tokenized URLs, blog and post titles, anchor text, blog homepage content, and post (entry) content
  - Feature selection using Fisher linear discriminant analysis
    - Avoid over-fitting
Detection performance

- Temporal features alone (R) out-perform the best 32 content features (base-32)
- Baseline and temporal features jointly work well
Summary

- Splog - new and serious problem in the blogosphere
- Splog detection is different from web spam detection
- Unique features of blogs/splogs
  - Content, temporal and link