

# Fast Splog Filtering: URL Language Model Segmentation

Franco Salvetti<sup>†\*</sup>

franco.salvetti@colorado.edu

Nicolas Nicolov<sup>\*</sup>

nicolas@umbrialistens.com

<sup>†</sup> Dept. of Computer Science, Univ. of Colorado at Boulder, Boulder, CO 80309-0430, U.S.A.

<sup>\*</sup> Umbria, Inc., 1655 Walnut St., Boulder CO 80302, U.S.A.

# Bird's Eye View

- Big Goal:** Analyze the blogosphere.
- BIG Obstacle:** Spam.
- Existing Solutions:** Slow, inaccurate.
- Insight:** Spammers glue words in URL—segment.
- New Approach:** Spam filter based on segmented URLs.

# Weblogs

- Personal journal on the web. Weblogs express as many different subjects and opinions. Some blogs are highly influential and have enormous readership; others are primarily intended for a close circle of family and friends.
- 42.1 million sites and 2.5 billion links.
- Blogosphere doubles in size every 5 months.
- Useful for: marketing intelligence; trend discovery; opinion tracking.

# BIG Problem: Spam Weblogs (Splogs)

Weblogs that promote affiliated web sites.

- There are lots of them (some bloghosts allow automatic blog creation).
- Slow processing.
- Storage ramifications.
- Skew results of analysis.

# Insight

- Sploggers want to communicate a concept.
- They do so in the URL.
- To avoid detection sploggers glue words: `dailyfreeipod`.

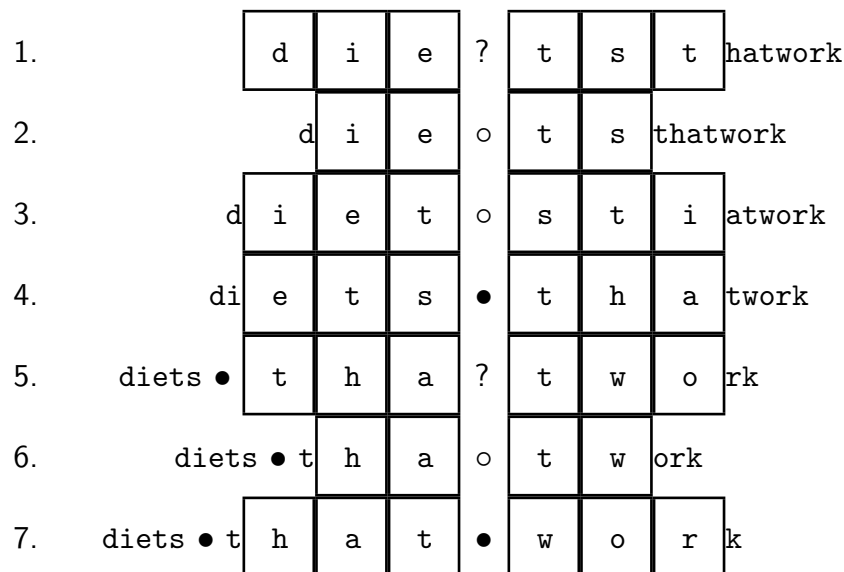
**We can segment the URLs.** Then classify as “usual”.

# Segmentation: Symmetric Sliding Window

'?' left and right  $n$ -grams not encountered in training;

'o' kept together;

'●' indicates a break.



# Example (Correct) Segmentations

cash • for • your • house

unlimited • pet • supplies

jim • and • body • fat

weight • loss • product • info

kick • the • boy • and • run

bringing • back • the • past

food • for • your • speakers

# Classification

Token sequence  $T = \langle t_1, \dots, t_n \rangle$  – the segmented URL

The class  $\hat{c} \in C = \{\text{spam}, \text{good}\}$ . Naïve Bayes:

$$\begin{aligned}\hat{c} &= \arg \max_{c \in C} P(c|T) = \arg \max_{c \in C} \frac{P(c) \cdot P(T|c)}{P(T)} \\ &= \arg \max_{c \in C} P(c) \cdot P(T|c) \\ &= \arg \max_{c \in C} P(c) \cdot \prod_{i=1}^n P(t_i|c)\end{aligned}$$



# Data

`http:// free - webcam - girl . net`

`http:// download|ringtone|for|tmobile . power|play - ringtones . com`

`http:// business|opportunity|money|work|at|home . coolblogstuff . com`

`http:// honda|parts|unlimited . freeblogsearch . info`

`http:// generator|deals . com / blog`

`http:// www . we|make|beer . com / blog`

`http:// the|canadian|sentinel . blogspot . com`

`http:// do|nt|mess|with|taxes . typepad . com`

10K spam blogs; 10K “good” blogs; 1K dev & 1K test URLs.

## Extra Segmentation

# of splits	# spam URLs	# good URLs
1	2,235	2,274
2	868	459
3	223	46
4	77	7
5	2	1
6	4	1
8	3	—
Total	3,412	2,788

# Classification Results

accuracy	78%
prec. spam	82%
rec. spam	71%
f-meas spam	76%
prec. good	74%
rec. good	84%
f-meas good	79%

# Human Performance

	$\mu$	$\sigma$
accuracy	76%	6.71
prec. spam	83%	7.57
rec. spam	65%	6.35
f-meas spam	73%	7.57
prec. good	71%	6.35
rec. good	87%	6.39
f-meas good	78%	6.08

Simple segmenter, simple classifier—beats humans!

# Ongoing Work

- Issues:
  - Plurals: 'girl●s' vs. 'girls'
  - Past tense: 'dedicate●d' vs. 'dedicated'
- Exploring:
  - Multi-break segmenters.
  - Analyzing outgoing links.
  - Better classifiers: Support Vector Machines (SVM); Robust Risk Minimization (RRM); Maximum Entropy, ...

# References

- Andras, T.S., A.Benczúr, K.Csalogány & M.Uher. 2005. "SpamRank—Fully Automatic Link Spam Detection". *WWW'05*.
- Delany, S.J., P.Cunningham , A.Tsymbal , L.Coyle 2005. "A Case-based Technique for Tracking Concept Drift in Spam Filtering". *Knowledge-Based Systems*, 18:2-3.187-195, Elsevier.
- Gao, J., M.Li & C.-N.Huang. "Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach". *Computational Linguistics* 31:4.531-574.
- Gyöngyi, Z., H.Garcia-Molina & J.Pedersen. 2004. "Combating Web Spam with TrustRank". *VLDB*.
- Hurst, M. 2005. "Deriving Marketing Intelligence from Online Discussion". *KDD'05*, 419-428.
- Kan, M.-Y. 2004. "Web Page Classification without the Web Page". *WWW'04*.
- Kan, M.-Y. & H.O.N.Thi. 2005. "Fast Webpage Classification Using URL Features". *CIKM'05*, 325-326.
- Kolari, P., T.Finin & A.Joshi. 2006. "SVMs for the Blogosphere: Blog Identification and Splog Detection". *Computational Approaches to Analyzing Weblogs*, 92-99.
- Mishne, G., D.Carmel & R.Lempel. 2005. Blocking Blog Spam with Language Model Disagreement. *AIRWeb'05*.
- SURBL. 2006. "Surbl - Spam Uri Realtime Blocklists". <http://www.surbl.org>.
- Technorati. 2006. "State of the Blogosphere February 2006 part 1: on Blogosphere Growth, Feb'06".