Blog Track Open Task: Spam Blog Classification*

Pranam Kolari, Tim Finin, Akshay Java
Anupam Joshi and Justin Martineau
University of Maryland Baltimore County
Baltimore, MD 21250, USA
{kolari1, aks1, finin, joshi, jm1}@umbc.edu

James Mayfield
Johns Hopkins University Applied Physics Laboratory
Laurel, MD 20723, USA
james.mayfield@jhuapl.edu

Abstract

Spam blogs or Splogs are blogs with either auto-generated or plagiarized content created for the sole purpose of hosting ads, promoting affiliate sites and getting new pages indexed. Splogs now rival generic web spam and e-mail spam, presenting a major problem to analytics on the blogosphere from basic search and indexing, to opinion, community, influence and correlation detection. This open task submission details how splogs impact *Opinion Identification*, and proposes an approach to assessment and evaluation for a Spam Blog Classification task in 2007.

Introduction

Blogs are radically changing the face of communications on the Internet. Beyond publishing content, blogs enable users to engage in conversation and form tight knit communities, constituting a highly influential social medium on the Web. The personal nature of blogs coupled with the ability to index and analyze them almost instantaneously present new opportunities for social pulse analysis launching blog search, and buzz/influence tracking systems to the forefront, with applications ranging from market research and public relations to social psychology and political science.

Though the potential benefit from blogs to both producers and consumers is unquestionable, their promotion is hindered by a disturbing problem that now afflicts the blogosphere, that of *Spam*, which now inundates the blogosphere. Spam blogs or splogs refer to blogs created for the sole purpose of hosting ads, promoting page rank of affiliates and getting new content indexed. The problem is amplified by the fact that content is either auto-generated or plagiarized from other sources.

The scale of the problem has now assumed shocking proportions. Splogs generated using readily available splog creation software (Finin 2006), now overwhelm the blogosphere both at ping servers, and at systems that index and analyze blogs, as reported by blog search and analysis engines (Umbria 2005; Cuban 2005), popular bloggers (Rubel 2005), and more recently through a formal study by us (Kolari, Java, & Finin 2006). This analysis makes some disturbing conclusions on spam faced by ping servers, systems that relay pings from freshly updated blogs to blog search engines. Approximately 75% of such pings are received from splogs. Downstream at blog search engines these numbers are estimated to be around 20% (Kolari *et al.* 2006).

Splogs are a major concern to blog publishing platforms and hosting services, they skew results for blog analytics and Web advertising campaigns, making blog based analysis less effective and less trustworthy. This growing problem motivates our proposal on a spam blog classification task in 2007.

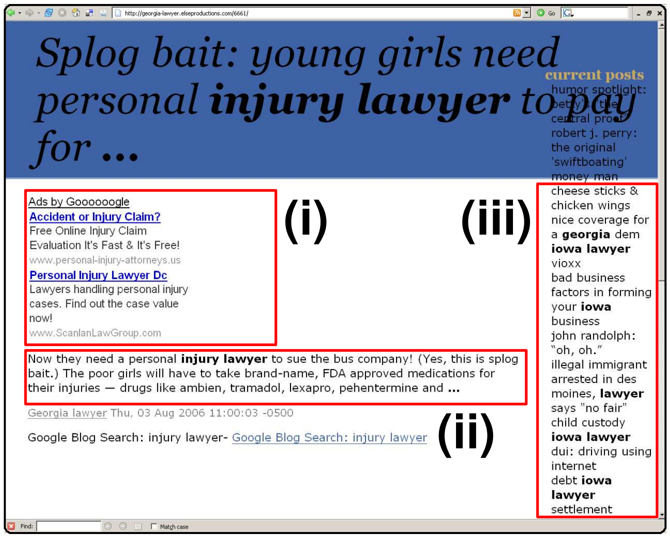
The rest of this open task submissions is structured as follows. The next section shows the basic anatomy of a spam blog, defines spam blog detection and places it in the context of spam in the Internet. We then survey the area of spam blog detection, techniques that work, along with showing the nature of the problem in the TREC Blog track 2007 and its affect on the primary task of opinion detection. We then present our proposal on assessing spam, and discuss evaluation approaches. Finally we discuss the proposed input and output formats before concluding the proposal.

Splog Detection Problem

Figure 1 shows a post from a splog, obtained by querying the index of a popular blog search engine. As shown, it (i) displays ads in high paying contexts, (ii) features content plagiarized from other blogs, and (iii) hosts hyperlinks that create link farms. Scores of such pages now pollute the blogosphere, with new ones springing up every moment. Though splogs continue to be a problem for web search engines, they present a new set of challenges for blog analytics. We discuss them in detail in the rest of this section.

In the classical web graph model $G(X, E)$, the set of nodes X represent web pages, and the set of edges E stand for hyper-links between these pages. In contrast, blog search engines treat the Web using a slightly more intricate and tuned model, $G([B, N, W], E)$, where $X = B \cup N \cup W$. The membership of nodes in this web-graph is in either of B, N or W, where B is the set of all pages (permalinks) from blogs, N is the set of all pages from news-sources (edited content), and W is the set representing the rest of the Web. Splog detection is a classification problem within the blogosphere subset, B. Typically, the result of such a classification leads to disjoint subsets B_A, B_S, B_U where

[*] Partial support was provided by an IBM Fellowship and by NSF awards NSF-ITR-IIS00326460 and NSF-ITR-IDM-0219649



Splog bait: young girls need personal **injury lawyer** to pay for ...

current posts
humor spotlight: betty's 'the central proof'
robert j. perry: the original 'swiftboating' money man
cheese sticks & chicken wings
nice coverage for a **georgia** dem
iowa lawyer
vioxx
bad business factors in forming your **iowa** business
john randolph: "oh, oh."
illegal immigrant arrested in des moines, **lawyer** says "no fair"
child custody
iowa lawyer
dui: driving using internet
debt **iowa lawyer** settlement

Ads by Goooooogle
Accident or Injury Claim?
Free Online Injury Claim Evaluation It's Fast & It's Free!
www.personal-injury-attorneys.us
Personal Injury Lawyer Dc
Lawyers handling personal injury cases. Find out the case value now!
www.ScanlanLawGroup.com

Now they need a personal **injury lawyer** to sue the bus company! (Yes, this is splog bait.) The poor girls will have to take brand-name, FDA approved medications for their injuries — drugs like ambien, tramadol, lexapro, pehentermine and **...**

Georgia lawyer Thu, 03 Aug 2006 11:00:03 -0500

Google Blog Search: injury lawyer- Google Blog Search: injury lawyer

Figure 1: A typical splog, plagiarizes content (ii), promotes other spam pages (iii), and (i) hosts high paying contextual advertisements

B_A represents all authentic content, B_S represents content from splogs and B_U represents those blog pages for which a judgment of authenticity or spam has not yet been made.

Using generic logistic regression the splog detection problem for any node $x \in B$, can be expressed as:

$$P(x \in B_S/O(x)); P(x \in B_S/L(x))$$

$$P(x \in B_S/L(x), O(x))$$

where $x \in B$, $O(x)$ represents local features, and $L(x)$ represents the link features.

On cursory investigation, this might still appear to be a classical web classification problem, and specifically a variation on the web spam problem (Gyöngyi & Garcia-Molina 2005) addressed by TrustRank (Gyöngyi, Garcia-Molina, & Pedersen 2004). However, the methodologies used by blog search engines and the nature of the blogosphere make this an interesting special case of web classification.

- **Search Engine Coverage.** Splogs are created with the intention of gaming web search engines associated with high traffic referrals, and are typically part of link-farms that go beyond the blogosphere. Blog search engines however work on a much smaller subset of the Web, by employing preferential crawling and indexing towards the sets B and N. Crawling external pages is often impractical for blog search engines. Give this, any feasible solution should be based on a combination of local models and link-based models using pages within B.

- **Quicker Assessment.** In addition to reach, blog search engine are judged by how quickly they can index and analyze new content. Since users are uninterested in splogs but interested in new content, splogs must be eliminated quickly. This emphasis on speed differentiates splog detection from classical web spam detection that is usually applied days after content creation.

- **Genre of Blog Content.** Most automated web spam detection techniques ignore local features in identifying spam pages, since content spans across almost all topics. The blogosphere, however, is a medium for specific genres like personal opinions and journals; and for unedited content, words appearing on a page can provide interesting cues for classification.

We next discuss some of the techniques that have been effective in detecting spam blogs.

Detecting Splogs

Over the past year we have been developing techniques to detect spam blogs. Our thrust has been on using local models, though we have also explored the use of link-based models.

Our work is based on a seed data set (Kolari, Finin, & Joshi 2006) of 700 positive (splogs) and 700 negative (authentic blog) labeled examples containing the entire HTML content of each blog home-page. All of the models are based on SVMs (Boser, Guyon, & Vapnik 1992), which are known to perform well in classification tasks (Joachims 1998). This paper only reports the highlights based on a linear kernel with top features chosen using mutual information. Interested readers are referred to (Kolari *et al.* 2006) for further details.

A blog's local features can be quite effective for splog detection. A *local feature* is one that is completely determined by the contents of a single web page, i.e. it does not require following links or consulting other data sources. A local model built using only these features can provide a quick assessment of the authenticity of blogs. We have experimented with many such models, and our results are summarized in Figure 2.

Words

To verify their utility, we created bag-of-words for the samples based on their textual content. To analyze discriminating features we used a simple approach of ordering features by weights assigned to them in the SVM model. It turns out that the model was built around features which the human eye would have typically overlooked. Blogs often contain content that expresses personal opinions, so words like "I", "We", "my", "what" appear commonly on authentic blog posts. To this effect, the bag-of-words model is built on an interesting "blog content genre". In general, such a content genre is not seen on the Web, which partly explains why spam detection using local textual content is less effective there.

Word N-Grams

An alternative methodology to using textual content for classification is the bag-of-word-N-Grams, where N adjacent words are used as a feature. We evaluated both bag-of-word-2-Grams and bag-of-word-3-Grams, which turned out

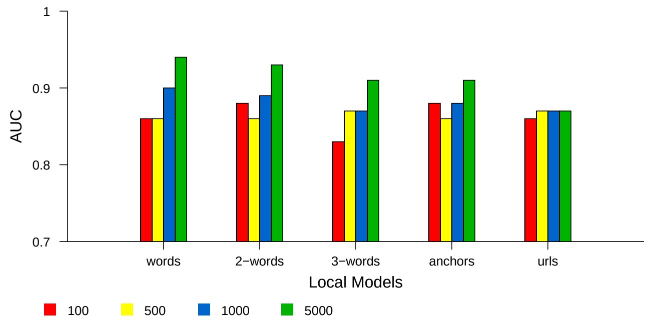


Figure 2: The performance of local models using different feature types and size for a linear kernel. Top features were ranked using mutual information.

to be almost as effective as bag-of-words. Interesting discriminative features were observed in this experiment. For instance, text like "comments-off" (comments are usually turned-off in splogs), "new-york" (a high paying advertising term), "in-uncategorized" (spammers do not bother to specify categories for blog posts) are features common to splogs, whereas text like "2-comments", "1-comment", "i-have", "to-my" were some features common to authentic blogs. Similar features ranked highly in the 3-word gram model.

Tokenized Anchors

Anchor text is the text that appears in an HTML link (i.e., between the `<a...>` and `</a>` tags.) We used a bag-of-anchors feature, where anchor text on a page, with multiple word anchors split into individual words, is used. Anchor text is frequently used for web page classification, but typically to classifying the target page rather than the one hosting the link. We observed that "comment" and "flickr" were among the highly ranked features for authentic blogs.

Tokenized URLs

Intuitively speaking, both local and outgoing URLs can be used as effective attributes for splog detection. This is motivated by the fact that many splogs point to the ".info" domain, whereas many authentic blogs point to well known websites like "flickr", "technorati" and "feedster" and strongly associated with blogs. We term these features as bag-of-urls, arrived at by tokenizing URLs using "/" and ".".

Global Models

A global model is one that uses some non-local features, i.e., features requiring data beyond the content of Web page under test. We have investigated the use of link distributions to see if splogs can be identified once they place themselves on the blog (web) hyper-link graph. The intuition is that that authentic blogs are very unlikely to link to splogs and that splogs frequently do link to other splogs. We have evaluated this approach by extending our seed dataset with in-links and out-links, to achieve AUC values of close to 0.85. Interested readers are referred to (Kolari *et al.* 2006) for further details.

Other Techniques

Some of the techniques used by update ping servers, blog publishing hosts and search engines are documented on the Web. We list them here for completeness.

- Update ping servers eliminate spings using ping frequency. Blogs (possibly splogs) that ping too frequently are blacklisted and their pings are discarded for a subsequent time window.
- In (Salvetti & Nicolov 2006) a URL tokenization approach is suggested, based on just using the URL of the updated blog, tokenized to identify words. The authors claim to achieve precision values as high as 93.3% and recall of 50.9%. Though recall is low, the advantage of this technique comes from speed; splog filtering is achieved without fetching content from the blog in question.
- Extensions on blog comment spam model seem to be used in certain cases. Akismet[1], a blog comment spam plug-in claims to detect splogs using comment spam models.
- URL/IP blacklists created by splog fighting efforts like http://splogspot.com, http://fightsplog.net are also in use by some systems.

TREC Blog Track 2006

TREC Blog Track 2006 asked participants to implement and evaluate a system for "opinion retrieval" from blog posts. Specifically, the task was defined as follows: build a system that will take a query string describing a topic, e.g., "March of the Penguins", and return a ranked list of blog posts that express an opinion, positive or negative, about the topic. For evaluation, NIST provided a dataset of over three million blogs drawn from about 80 thousand blogs. Participants built and trained their systems to work on this dataset. Contestants do an automatic evaluation by downloading and running, without further modification to their systems, a set of fifty test queries. The results are currently being evaluated by NIST and will be available in November.

We studied the impact of splogs during our own participation in TREC (Java *et al.* 2006), and report them here. Our analysis is based on a splog detection technique that works on a blog home-page using word features. We argue that splogs significantly impact blog analytics by showing how they affected the opinion retrieval task, and more generally query relevance.

Impact of Splogs

In order to make the challenge realistic NIST explicitly included 17,969 feeds from splogs, contributing to 15.8% of the documents (Macdonald & Ounis 2006). There were 83,307 distinct homepage URLs present in the collection, of which 81,014 could be processed. The collection contained a total of 3,214,727 permalinks from all these blogs. Our automated splog filter identified 13,542 splogs. This accounts

[1] http://akismet.com

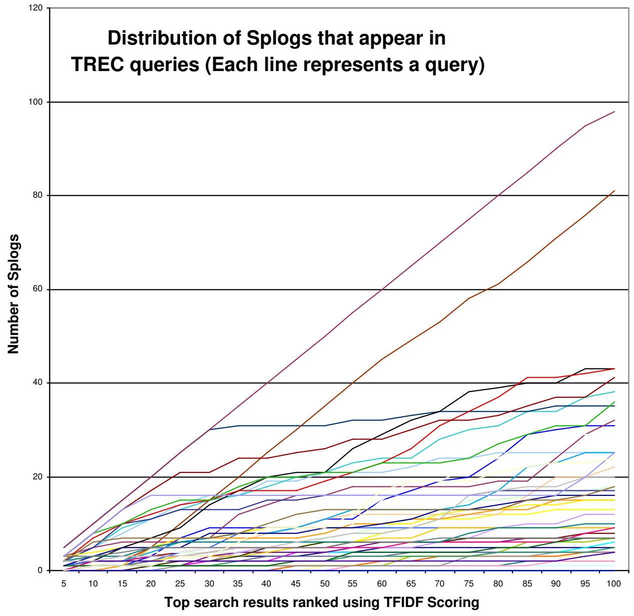


Figure 3: The number of splogs in the top x results for 50 TREC queries. Top splog queries include "cholesterol" and "hybrid cars"

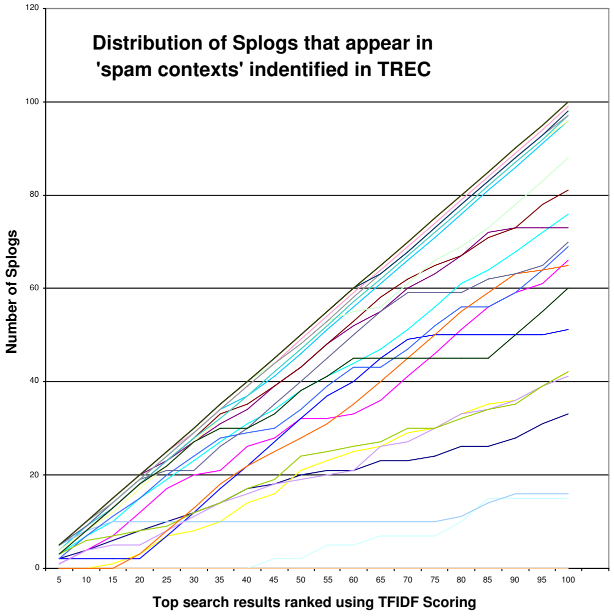


Figure 4: The number of splogs in the top x results of the TREC collection for 28 highly spammed query terms. Top splog queries include 'pregnancy', 'insurance', 'discount'

for about 16% of the identified homepages. The total number of permalinks from these splogs is 543,086 or around 16% of the collection. While the actual list of splogs is not available for comparison until the completion of the TREC open task deadline, the current estimate appears close. To prevent the possibility of splogs skewing our results permalinks associated with splogs were not indexed in the system we built for TREC 2006.

To keep the analysis generic, we evaluate the influence of splogs in the context of search engine retrieval. Given a search query, we would like to estimate the impact splogs have on search result precision. Figure 3 shows the distribution of splogs across the 50 TREC queries. The number of splogs present varies across the queries since splogs are query (topic) dependent. For example, the topmost spammed query terms were 'cholesterol' and 'hybrid cars'. Such queries attract a high paying advertisement market, which splogs exploit.

The description of the TREC collection (Macdonald & Ounis 2006) provides an analysis of posts from splogs that were added to the collection. Top informative terms include 'insurance', 'weight', 'credit' and such. Figure 4 shows the distribution of splogs identified by our system across such spam terms. In stark contrast from Figure 3 there is a very higher percentage of splogs in the top 100 results.

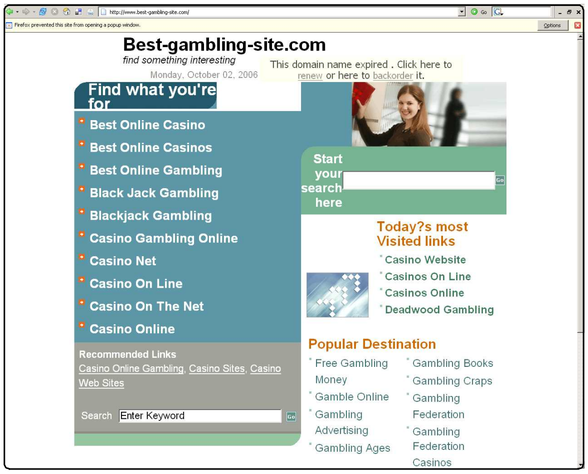
Splog Task Assessment

We approach splog detection at the blog home-page level, and then propagate to all posts (permalinks) hosted by the blog. Although this seems to work well in practice, is in use by a blog search engine and a partner academic institution, we propose a more structured approach for TREC.

Inspired by e-mail spam detection, we argue that permalinks (individual posts) be treated as atomic entities for assessment (and hence evaluation) in the splog detection task, irrespective of whether splogs are detected at the post or blog home-page level. Independent of IP blacklists, blacklisted e-mail addresses (analogous to blog home-pages) and proxies, e-mail spam detection techniques are evaluated on a per-mail basis. For a splog detector permalinks could be treated analogously to emails received at an address, providing intuition and structure to the task. This also aligns with how blog search engines tap into blog updates, making on the fly decisions about indexing new posts.

We propose a categorization scheme (possibly overlapping) for spam permalinks based on a web spam taxonomy scheme proposed by (Gyöngyi & Garcia-Molina 2005), and our experience dealing with spam blogs.

- **non-blog** pages attempt to associate themselves with the blogosphere to exploit increased search engine exposure. Non-blogs usually infiltrate the blogosphere through unjustified pings at update ping servers. See figure 5.
- **keyword-stuffing** targets tfidf (Salton & Buckley 1988) based relevance measures used by search engines. Spam posts repeat target keywords (query terms) multiple times on their pages. See figure 6.
- **post-stitching** is used to automatically generate content,


http://www.best-gambling-site.com/
Firefox prevented this site from opening a popup window.
Options

Best-gambling-site.com

find something interesting

Monday, October 02, 2006

This domain name expired . Click here to renew or here to backorder it.

Find what you're for

- Best Online Casino
- Best Online Casinos
- Best Online Gambling
- Black Jack Gambling
- Blackjack Gambling
- Casino Gambling Online
- Casino Net
- Casino On Line
- Casino On The Net
- Casino Online

Recommended Links

Casino Online Gambling, Casino Sites, Casino Web Sites

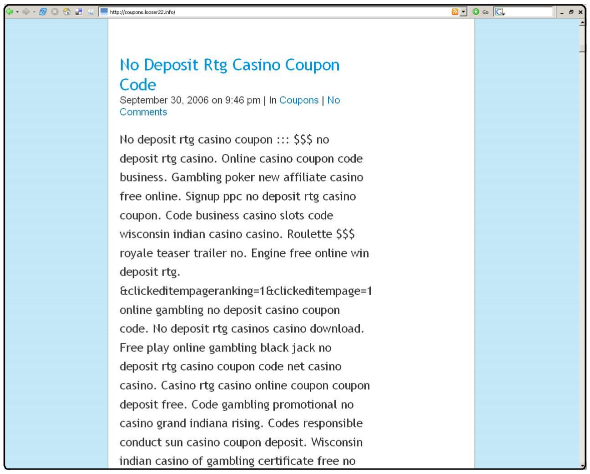
Today?s most Visited links

- Casino Website
- Casinos On Line
- Casinos Online
- Deadwood Gambling

Popular Destination

- Free Gambling Money
- Gamble Online
- Gambling Advertising
- Gambling Ages
- Gambling Books
- Gambling Craps
- Gambling Federation
- Gambling Federation Casinos

Figure 5: A non-blog page that pings weblogs.com to exploit the increased search engine exposure of the blogosphere.

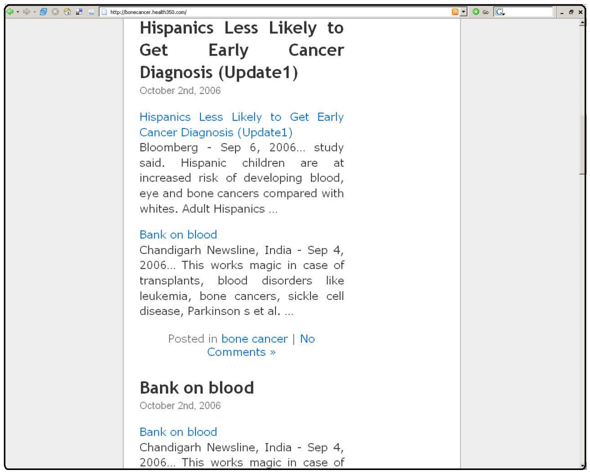


No Deposit Rtg Casino Coupon Code

September 30, 2006 on 9:46 pm | In Coupons | No Comments

No deposit rtg casino coupon ::: $$$ no deposit rtg casino. Online casino coupon code business. Gambling poker new affiliate casino free online. Signup ppc no deposit rtg casino coupon. Code business casino slots code wisconsin indian casino casino. Roulette $$$ royale teaser trailer no. Engine free online win deposit rtg. &clickeditempageranking=1&clickeditempage=1 online gambling no deposit casino coupon code. No deposit rtg casinos casino download. Free play online gambling black jack no deposit rtg casino coupon code net casino casino. Casino rtg casino online coupon coupon deposit free. Code gambling promotional no casino grand indiana rising. Codes responsible conduct sun casino coupon deposit. Wisconsin indian casino of gambling certificate free no

Figure 6: A spam post using keyword stuffing technique to rank highly for “coupon code”



Hispanics Less Likely to Get Early Cancer Diagnosis (Update1)

October 2nd, 2006

Hispanics Less Likely to Get Early Cancer Diagnosis (Update1)

Bloomberg - Sep 6, 2006... study said. Hispanic children are at increased risk of developing blood, eye and bone cancers compared with whites. Adult Hispanics ...

Bank on blood

Chandigarh Newsline, India - Sep 4, 2006... This works magic in case of transplants, blood disorders like leukemia, bone cancers, sickle cell disease, Parkinson s et al. ...

Posted in bone cancer | No Comments »

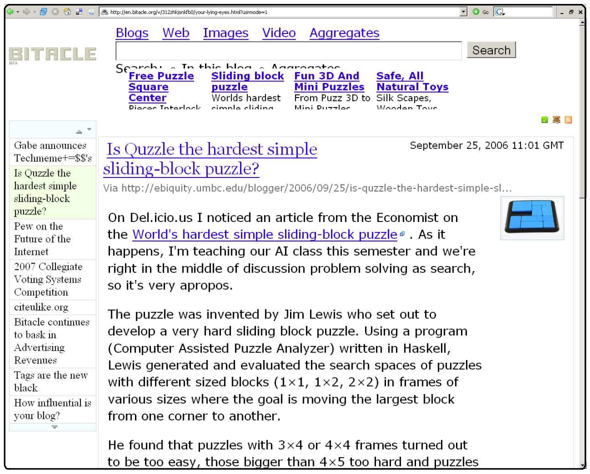
Bank on blood

October 2nd, 2006

Bank on blood

Chandigarh Newsline, India - Sep 4, 2006... This works magic in case of

Figure 7: A spam post that stitches together excerpts in a highly profitable advertising context.


http://en.bitacle.org/v/312zhkjsnkfb0/your-lying-eyes.html?usrmode=1
BITACLE
BETA
Blogs
Web
Images
Video
Aggregates
Search
Search: In this blog Aggregates

Free Puzzle Square Center
Pieces Interlock

Sliding block puzzle
Worlds hardest simple sliding

Fun 3D And Mini Puzzles
From Puzz 3D to Mini Puzzles

Safe, All Natural Toys
Silk Scapes, Wooden Toys

Gabe announces Techmeme+=$$'s

Is Quzzle the hardest simple sliding-block puzzle?

Pew on the Future of the Internet

2007 Collegiate Voting Systems Competition

citeulike.org

Bitacle continues to bask in Advertising Revenues

Tags are the new black

How influential is your blog?

Is Quzzle the hardest simple sliding-block puzzle?

September 25, 2006 11:01 GMT

Via http://ebiquity.umbc.edu/blogger/2006/09/25/is-quzzle-the-hardest-simple-sl...

On Del.icio.us I noticed an article from the Economist on the World's hardest simple sliding-block puzzle . As it happens, I'm teaching our AI class this semester and we're right in the middle of discussion problem solving as search, so it's very apropos.

The puzzle was invented by Jim Lewis who set out to develop a very hard sliding block puzzle. Using a program (Computer Assisted Puzzle Analyzer) written in Haskell, Lewis generated and evaluated the search spaces of puzzles with different sized blocks (1×1, 1×2, 2×2) in frames of various sizes where the goal is moving the largest block from one corner to another.

He found that puzzles with 3×4 or 4×4 frames turned out to be too easy, those bigger than 4×5 too hard and puzzles

Figure 8: A case of full content plagiarism without author consent by a blog re-publisher.

by combining excerpts from plagiarized posts or news stories, in high paying advertising contexts. See figure 7.

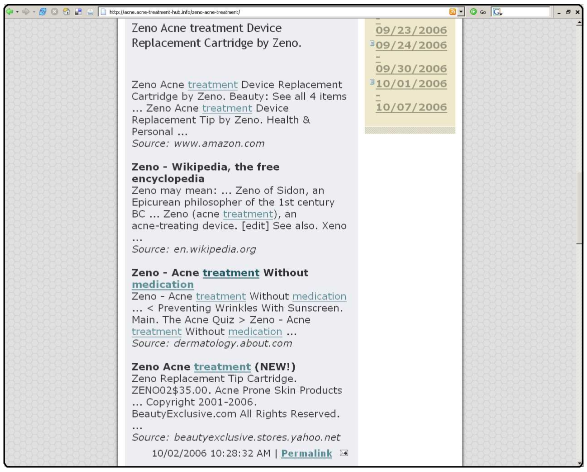
- **post-plagiarism** is full content theft from blogs or news pages. The recent debate surrounding Bitacle[2] is one such example. See figure 8.
- **post-weaving** is used to conceal excessive links to affiliate sites by copying entire posts from other blogs and weaving in hyperlinks promoting affiliates. See figure 9.
- **link-spam** is an attempt to artificially inflate page rank or get new pages indexed, using link dumps that contribute to a link-farm. Note that post-weaving can be considered a form of link-spam. See figure 10.
- **other-techniques** group techniques that appear in none of the classes above - common ones being page redirection and cloaking.

The motivation behind a categorization is that different detection models will have to be developed based on the category of spam to be detected. For instance, our own word based model works well for the keyword-stuffing and post-stitching category. We believe that such an explicit categorization will encourage the consideration of all aspects of spam blogs by task participants.

Approach

Our proposed assessment values and their interpretation is shown in table 1. Assessment is done to permalinks independent of query term or context.“-1” score represents a non-judgment, and is similar in semantics to its use in Blog Track 2006 opinion task. “0” represents an authentic blog post, and the rest of the scores are used to identify a spam post and its category.

[2] http://bitacle.org



Zeno Acne treatment Device Replacement Cartridge by Zeno.

Zeno Acne treatment Device Replacement Cartridge by Zeno. Beauty: See all 4 items ... Zeno Acne treatment Device Replacement Tip by Zeno. Health & Personal ...
Source: www.amazon.com

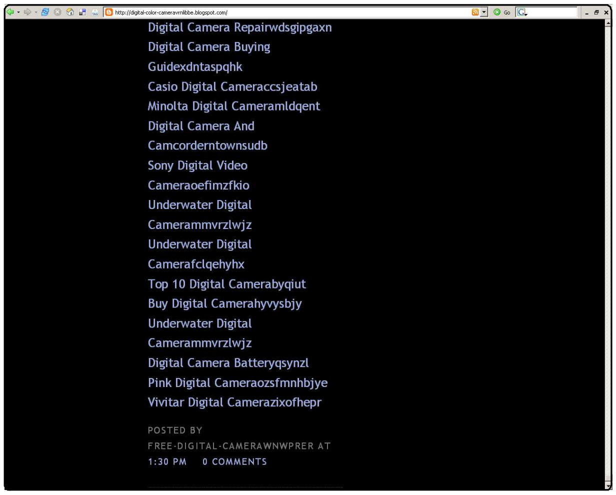
Zeno - Wikipedia, the free encyclopedia
Zeno may mean: ... Zeno of Sidon, an Epicurean philosopher of the 1st century BC ... Zeno (acne treatment), an acne-treating device. [edit] See also. Xeno ...
Source: en.wikipedia.org

Zeno - Acne treatment Without medication
Zeno - Acne treatment Without medication ... < Preventing Wrinkles With Sunscreen. Main. The Acne Quiz > Zeno - Acne treatment Without medication ...
Source: dermatology.about.com

Zeno Acne treatment (NEW!)
Zeno Replacement Tip Cartridge. ZENO02$35.00. Acne Prone Skin Products ... Copyright 2001-2006. BeautyExclusive.com All Rights Reserved. ...
Source: beautyexclusive.stores.yahoo.net

10/02/2006 10:28:32 AM | **Permalink**

Figure 9: Spam post weaving hyperlinks to affiliate sites for "acne treatment".



Digital Camera Repairwdsgipgaxn
Digital Camera Buying
Guidexdntaspqhk
Casio Digital Cameraccsjeatab
Minolta Digital Cameramldqent
Digital Camera And
Camcorderntownsudb
Sony Digital Video
Cameraoefimzfkio
Underwater Digital
Camerammvrzlwjz
Underwater Digital
Camerafclqehyhx
Top 10 Digital Camerabyqiut
Buy Digital Camerahyvysbjy
Underwater Digital
Camerammvrzlwjz
Digital Camera Batteryqsynzl
Pink Digital Cameraozsfmnhbjye
Vivitar Digital Camerazixofhepr

POSTED BY
FREE-DIGITAL-CAMERAWNWPRER AT
1:30 PM 0 COMMENTS

Figure 10: A spam post contributing to a link-farm involving other blogs.

Though assessments are attached to permalinks, the assessor labels blog home-pages which can then be propagated down to all permalinks from the blog. Based on our experience labeling spam blogs, each assessment takes around 1-2 minutes, as measured from the time the page was accessed to the time the assessment was entered.

Splog Task Evaluation

Assuming the existence of a TREC Collection similar to the collection of 2006 spanning n days, we propose that the dataset be divided into two subsets. The first subset, from here on referred to as D_{base} will span the first $(n - x)$ days of the collection and the second, referred to as D_{test} will span the last x days of the collection. The exact value of x will be decided collectively and could be one, two or more days.

$Score$	Interpretation
-1	Not judged
0	Authentic blog post
1	non-blog
2	keyword-stuffing
3	post-stitching
4	post-plagiarism
5	post-weaving
6	link-spam
7	other-techniques

Table 1: Proposed assessment scores for spam blog classification

D_{base} will be released at task announcement for participants to train and build their splog detection models. D_{test} will be released subsequently along with a input (test) file to the spam blog detector. Unlike TREC Blog 2006 where systems were judged by 50 independent topics (queries), the proposed task will be judged based on 50 independent sets of permalinks (sampled from D_{test}). The cardinality of each such set will be arrived at through further discussions. We believe this is a good model for what blog and feed search engines have to do i.e. make judgment on newly created posts based on knowledge gathered while indexing earlier posts, observed attributes of blogs vs. splogs, and models they built around them.

Spam blog detectors developed by participants will rank the set of permalinks based on an estimated "sploggieness". The overall evaluation of systems will be done just on this ranking, but the category data will allow participants to see where their systems were strong and weak, to informally compare across participants, and will serve as feedback for overall improvement of the quality of the blogosphere.

Dataset Creation

The approaches followed in the creation of the TREC 2006 Blog Collection is detailed in (Macdonald & Ounis 2006). In addition to permalinks, blog home-pages and feeds were also part of the collection, cached regularly polling a static list of blogs (and splogs) over a period of 77 days. One key component missing in this collection (and important) for a spam blog classification task is the dynamic aspect of newly created blogs and their posts; splogs are transient and short-lived.

To avoid replication of data collection efforts, an approach to create collections for multiple tasks together can be used. To overcome the problem[3] noted in (Macdonald & Ounis 2006) a ping server with better coverage[4], or multiple ping servers together can be first employed to tap into updates in the blogosphere. As a next step two collections could be created from the updates - (i) posts from all blog updates, and (ii) posts that intersect with the static list of authoritative bloggers used in TREC Blog Collection 2006. The first col-

[3] http://pubsub.com served pings for only 37% of the blogs in TREC Blog Collection 2006

[4] http://blo.gs can be employed over http://pubsub.com

lection can then be employed for spam blog classification, and the second for tasks around blog analytics.

Input/Output Formats

We propose the following formats for input and output files.

Input Format

The input file consists of 50 independent sets of permalinks. The association of a permalinks with home-page, syndication feed, and post time-stamp is also specified with the input. Track participants can use any of them or their combination, but make explicit which fields were used.

```
<set>
<num>...</num>
<test>
<permalink>
<url>...</url>
<homepage>...</homepage>
<feed>...</feed>
<when>... </when>
</permalink>
<permalink>
...
</permalink>
...
</test>
</set>
<set>
...
</set>
```

Output Format

The output format from a TREC run will be similar to the format used in the TREC Blog Track 2006 on Opinion Identification. Permalinks in each of the sets are to be ranked based on "sploggineess" score.

set Q0 docno rank prob runtag

where *set* is the input permalink set, *Q0* is literal "Q0", *docno* is the permalink identifier, *rank* is the final rank returned by the system, *prob* is the probability associated with spam judgment and *runtag* is the run's identifier string. Participants will be judged on precision/recall across a combination of all categories of splogs.

Conclusion

In this open task submission we have proposed a spam blog classification task for TREC Blog Track 2007, argued why it forms an important part of blog analytics and surveyed existing techniques on eliminating them. We have also shown how it impacted the primary task of Blog Track 2006 and put forward assessment and evaluation for such a task to be adopted in TREC Blog Track 2007.

References

Boser, B. E.; Guyon, I. M.; and Vapnik, V. N. 1992. A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, 144–152. New York: ACM Press.

Cuban, M. 2005. A splog here, a splog there, pretty soon it ads up and we all lose. [Online; accessed 22-December-2005; `http://www.blogmaverick.com/entry/1234000870054492/`].

Finin, T. 2006. Splog software from hell. [Online; accessed 31-August-2006; `http://ebiquity.umbc.edu/blogger/splog-software-from-hell/`].

Gyöngyi, Z., and Garcia-Molina, H. 2005. Web spam taxonomy. In *First International Workshop on Adversarial Information Retrieval on the Web*.

Gyöngyi, Z.; Garcia-Molina, H.; and Pedersen, J. 2004. Combating web spam with TrustRank. In *Proceedings of the 30th International Conference on Very Large Databases*, 576–587. Morgan Kaufmann.

Java, A.; Kolari, P.; Finin, T.; Mayfield, J.; Joshi, A.; and Martineau, J. 2006. The UMBC/JHU blogvox system. In *Proceedings of the Fifteenth Text Retrieval Conference*.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, 137–142. London, UK: Springer-Verlag.

Kolari, P.; Java, A.; Finin, T.; Oates, T.; and Joshi, A. 2006. Detecting Spam blogs: A machine learning approach. Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006).

Kolari, P.; Finin, T.; and Joshi, A. 2006. SVMs for the blogosphere: Blog identification and splog detection. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*.

Kolari, P.; Java, A.; and Finin, T. 2006. Characterizing the splogosphere. In *WWW 2006, 3rd Annual Workshop on the Webloggging Ecosystem: Aggregation, Analysis and Dynamics*.

Macdonald, C., and Ounis, I. 2006. The trec blogs06 collection: Creating and analyzing a blog test collection. Department of Computer Science, University of Glasgow Tech Report TR-2006-224.

Rubel, S. 2005. Blog content theft. [Online; `http://www.micropersuasion.com/2005/12/blog_content_th.html`].

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.* 24(5):513–523.

Salvetti, F., and Nicolov, N. 2006. Weblog classification for fast splog filtering: A url language model segmentation approach. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 137–140. New York City, USA: Association for Computational Linguistics.

Umbria. 2005. Spam in the blogosphere. [Online; `http://www.umbrialistens.com/consumer/showWhitePaper`].