Is Spam an Issue for Opinionated Blog Post Search?

Craig Macdonald, Iadh Ounis
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ, UK
{craigm,ounis}@dcs.gla.ac.uk

Ian Soboroff
NIST
Gaithersburg, MD, USA
ian.soboroff@nist.gov

ABSTRACT

In opinion-finding, the retrieval system is tasked with retrieving not just relevant documents, but those that also express an opinion towards the query target entity. This task has been studied in the context of the blogosphere by groups participating in the 2006-2008 TREC Blog tracks. Spam blogs (splogs) are thought to be a problem on the blogosphere. In this paper, we investigate the extent to which spam has affected the participating groups' retrieval systems over the three years of the TREC Blog track opinion-finding task. Our results show that spam can be an issue, with most systems retrieving some spam for every topic. However, removing spam from the rankings does not markedly change the relative performance of opinion-finding approaches.

Categories & Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

General Terms: Performance, Experimentation

Keywords: Blogs, opinion-finding, spam, splogs

1. INTRODUCTION

The Blog track at the annual Text REtrieval Conference (TREC) was initiated in 2006 to study the information-seeking behaviour in the blogosphere [3]. A key feature that distinguishes blog contents from the factual content used in other TREC tasks is their subjective nature. Many blog queries are person names, both celebrities and unknown, and the underlying user information needs seem to be of an opinion, or perspective-finding nature, rather than fact-finding. Indeed, opinion-finding has formed one of the central tasks investigated by the TREC Blog track since 2006 [3]. In this task, the information retrieval (IR) systems of participating groups are tasked with locating blog posts that express an opinion about the target entity, e.g. a name of a person, location or organisation.

Adversarial issues affect all forms of Web search, and the blogosphere is no exception. The ease by which blog posting can be automated has given rise to spam blogs (also known as splogs) [1]. Splogs may have plagiarised content from other blogs or news sources, in order to drive users to profitable context-based advertisements. Alternatively, false blogs with negligible content are created to realise a link farm intended to increase the search engine ranking of affiliated sites [1].

The TREC Blogs06 corpus has been used for the opinion-finding tasks in TREC 2006-2008. The corpus is comprised of blog post documents from 100,649 blogs, monitored over an 11 week period in late 2005 and early 2006 [4]. In this corpus, there is a proportion of assumed splogs that were injected into the corpus and crawled in the same manner as the normal blogs. The number of assumed splog feeds is 17,958 (17% of all feeds), which produced 509,137 posts (16%).

It is of note that the list of splogs was not released to groups participating in the opinion-finding tasks run as part of the TREC Blog track[1]. Hence, groups participating in TREC may have been affected by the presence of spam. In a previous study, we found that different topics were more or less affected by spam – e.g. health and transport topics were more likely to retrieve splog posts than topics on computers or shopping [2]. In contrast, this work analyses the effect that these spam blogs in the collection have had on the IR systems of the participating groups across the three years that the task ran (150 topics, 393 runs). Hence, it is the first comprehensive account of the implications of this spam on the absolute and relative effectiveness of various opinion-finding retrieval techniques.

2. METHODOLOGY & ANALYSIS

In this study, we ascertain how the splogs in the Blogs06 corpus affect the IR systems of participating TREC groups. For each year of the TREC Blog track opinion-finding task, 50 new query topics were selected from a commercial query log by TREC assessors - numbered 851-900, 901-950 and 1001-1050 for years 2006, 2007 and 2008, respectively. Participating groups run their IR systems on these topics, and submit the top 1,000 retrieved posts for each topic (known as a run). For TREC 2008, systems used submitted baseline runs (runs without opinion-finding features enabled) as the basis for their opinion-finding runs. This allows measurement of the relative improvement in opinion-finding performance compared to the corresponding baseline. Finally, the TREC 2008 submitted runs contained the retrieved documents for all 150 topics from all three years, allowing for the study into how the participating IR systems have evolved.

We believe that the presence of posts from splogs in a search engine ranking will annoy users, leading them to believe that the overall quality of the search engine is poor, with the consequent fall in usage, ad revenue, etc. Hence, based on the ground truth list of splogs in the corpus, we use standard IR measures to measure the presence of spam in the retrieved blog posts: Spam@R, the number of splog posts in the top R retrieved posts; Spam@all, the total number of splog posts in all of the 1,000 retrieved posts; Spam-MRR, the mean reciprocal rank that the first splog post retrieved for each query appears at; and finally Spam-MAP, the mean average precision where the splog posts are treated as rele-

[1] Now available from http://ir.dcs.gla.ac.uk/trec-blog

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

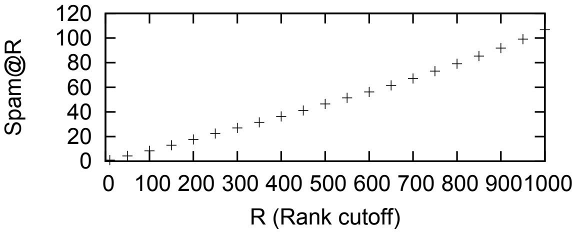

Figure 1: Mean number of splog posts retrieved by various rank cutoff points. 232 runs from 20 groups participating in TREC 2008, topics 1001-1050.

Topics	Run Year	Spam-MRR	Spam@10	Spam@all	Spam-MAP ×10⁻5
851-900	2006	0.1522	0.74	95.20	3.2
	2008	0.1169	0.48	93.68	3.6
901-950	2007	0.1823	0.84	102.44	6.8
	2008	0.1493	0.78	101.08	5.2
1001-1050	2008	0.1010	0.52	75.38	1.2

Table 1: Spam measures on the various topic sets and TREC years.

vant. For each of these measures, the higher the value, the more splog posts are being retrieved (undesirable). Note that because the splog posts are only occasionally retrieved, the absolute values for MAP are very low.

In the following, we analyse the systems that participated in the opinion-finding tasks that ran as part of the TREC 2006, 2007 and 2008 Blog tracks. Firstly, for the TREC 2008 runs on topics 1001-1050, Figure 1 shows the mean Spam@R for various R values. We can see that the number of spam documents increases linearly with rank, and that a marked number of these are retrieved in the first 100. Next, Table 1 reports the per-topic median spam measures on the three opinion-finding topic sets, showing how well the systems did in getting rid of spam (lower is better). The number of submitted runs were 57, 104 and 232[2] for years 2006, 2007 and 2008, respectively. From the results, we can see that the 2008 topics appear to be the least affected by spam (lowest Spam@10, Spam@all & Spam-MAP values). Moreover, for the topic sets that were reused in TREC 2008, the Spam@10 and Spam@all measures are reduced, suggesting that the systems in TREC 2008 were retrieving less splog posts than the TREC 2006/07 systems for the same topics.

Next, we investigate whether spam impacted on the retrieval effectiveness of the opinion-finding approaches. For TREC 2008, five submitted highly-performing baseline runs were selected by TREC and re-distributed to participants. The participants could try up to four opinion-finding approaches on top of each standard baseline [3]. However, on inspection, we find that all of these standard baselines were affected by splog posts, retrieving an average of 90-146 splog posts per 1000 posts, which is higher than the median retrieval systems reported in Table 1. Indeed, for topic 1035 (a health-related query), the baseline with highest opinion-finding retrieval performance (baseline4) also retrieved 444 splog posts out of 1000 posts.

We now investigate if any of the 21 opinion-finding approaches applied on all of the TREC 2008 five standard baselines would have benefited from splog post removal. Firstly, note that assessors (who were not asked to judge if a post was spam or not) may have judged splog posts as relevant – indeed, across the relevance assessments of all topics, 5.5% of all opinionated relevant posts were from splogs. Hence, we remove all splog posts from the relevance assessments. Next, from the baseline runs and the resulting opinion-finding approaches, we remove all splog posts (and denote this 'Without Spam'). In doing so, we assume that the opinion-finding techniques deployed by participants would have ranked the non-splog posts identically if the splog posts were or were not present in the corresponding input baseline run. Table 2 reports the mean performances of the best approaches by the top 3 groups out of the 21 approaches that used the 5 standard TREC 2008 baselines. The results are in terms of opinion-finding MAP and Δ opinion-finding MAP over the corresponding baseline, both with spam and when spam is removed from the runs. From Table 2, we see that while the relative ranking of groups' approaches appears unchanged, the overall effectiveness is increased. This shows that removing spam can increase the retrieval effectiveness of the opinion-finding approaches, if the spam is considered as non-relevant. Moreover, when spam is removed, user annoyance at spam should be reduced. Finally, if we correlate the relative ordering by opinion-finding MAP of all 21 opinion-finding approaches before and after spam removal, we obtain Kendall's $\tau = 0.962$, showing that there is very little difference in the relative ordering – indeed, the most effective approaches remain effective when spam is removed.

Group	Approach	With Spam		Without Spam	
		Mean MAP	Mean Δ MAP	Mean MAP	Mean Δ MAP
UIC IR	uicop1bl1r	0.3453	11.63%	0.3638	9.53%
KLE	B1PsgOpinAZN	0.3408	9.65%	0.3631	8.77%
UoGtr	uogOP1PrintL	0.3278	5.70%	0.3477	4.41%

Table 2: Mean performance across the top 3 approaches using all 5 standard TREC 2008 baselines, when spam is present and when removed. All splog posts are removed from the relevance assessments.

3. CONCLUSIONS

This paper has shown that spam is *currently* an issue in opinionated blog search. Our results lead us to infer that, in general, retrieval performance is benefited when spam is removed, and, moreover, user annoyance should be reduced. In particular, we find that, on average, for every 1,000 posts retrieved, the participating systems would retrieve 100 splog posts. This is a marked recall rate, and suggests that splog removal may be necessary. We also find that removal of spam did increase the effectiveness of the opinion-finding approaches. However, as it did not impact the relative ranking of the approaches, those approaches which are effective in the presence of spam are generally improved when spam is removed. In the future, we hope that by releasing the assumed spam ground truth in the Blogs06 dataset, more research and development on appropriate techniques for splogs identification can take place. With such a dataset, it should become possible for groups to investigate improving the absolute performance of their opinion-finding approaches in future TREC Blog track tasks, which will use the larger Blogs08 corpus with further possible spam.

4. REFERENCES

[1] P. Kolari, A. Java, and T. Finin. Characterizing the Splogosphere. In *Proceedings of 3rd WWE Workshop at WWW'06*, Edinburgh, UK, 2006.

[2] I. Ounis, C. Macdonald, and I. Soboroff. On the TREC Blog Track. In *Proceedings of ICWSM-2008,* Seattle, USA, 2008.

[3] I. Ounis, C. Macdonald, I. Soboroff. Overview of TREC-2008 Blog track. In *Proceedings of TREC-2008,* Gaithersburg, USA, 2009.

[4] C. Macdonald and I. Ounis. The TREC Blog06 Collection : Creating and Analysing a Blog Test Collection. *DCS Technical Report TR-2006-224*. Univ. of Glasgow. 2006. `http://www.dcs.gla.ac.uk/~craigm/publications/macdonald06creating.pdf`

[2] This total includes baseline runs.