Overview of the TREC-2006 Blog Track

Iadh Ounis, Maarten de Rijke, Craig Macdonald, Gilad Mishne, Ian Soboroff*

trecblog-organisers@dcs.gla.ac.uk

1 Introduction

The rise on the Internet of blogging, the creation of journal-like web page logs, has created a highly dynamic subset of the World Wide Web that evolves and responds to real-world events. Indeed, blogs (or weblogs) have recently emerged as a new grassroots publishing medium. The so-called blogosphere (the collection of blogs on the Internet) opens up several new interesting research areas.

Blogs have many interesting features: entries are added in chronological order, sometimes at a high volume. In addition, many blogs are created by their authors, not intended for any sizable audience, but purely as a mechanism for self-expression. Extremely accessible blog software has facilitated the act of blogging to a wide-ranging audience, their blogs reflecting their opinions, philosophies and emotions. Traditional media tends to focus on "heavy-hitting" blogs devoted to politics, punditry and technology. However, there are many different genres of blogs, some written around a specific topic, some covering several, and others talking about personal daily life [3].

The Blog track began this year, with the aim to explore the information seeking behaviour in the blogosphere. For this purpose, a new large-scale test collection, namely the TREC Blog06 collection, has been created. In the first pilot run of the track in 2006, we had two tasks, a main task (opinion retrieval) and an open task. The opinion retrieval task focuses on a specific aspect of blogs: the opinionated nature of many blogs. The second task was introduced to allow participants the opportunity to influence the determination of a suitable second task (for 2007) on other aspects of blogs, such as the temporal/event-related nature of many blogs, or the severity of spam in the blogosphere.

The remainder of this paper is structured as follows. Section 2 provides a short description of the newly created Blog06 test collection. Section 3 describes the opinion task, and provides an overview of the submitted runs of the participants. Section 4 describes the open task and the submitted proposals. We provide concluding remarks in Section 5.

2 Blog06 Test Collection

For the purposes of the TREC Blog track, there was a need to create a test collection of blog data. Such a collection should be a realistic snapshot of the blogosphere, containing enough blogs as to have recognisable properties of the blogosphere, and over a long enough time period that events should be recognisable. In addition, the collection should exhibit other properties of the blogosphere, such as splogs and comments spam. A new collection, called Blog06, was created by the University of Glasgow.

The collection included a selection of "top blogs" provided by Nielsen BuzzMetrics and supplemented by the University of Amsterdam. Moreover, a selection of blogs of genres accessible to the TREC assessors was

[*] Iadh Ounis and Craig Macdonald are affiliated to the University of Glasgow, UK; Maarten de Rijke and Gilad Mishne are affiliated to the University of Amsterdam, Netherlands; Ian Soboroff is affiliated to NIST, USA.

Quantity	Value
Number of Unique Blogs	100,649
RSS	62%
Atom	38%
First Feed Crawl	06/12/2005
Last Feed Crawl	21/02/2006
Number of Feeds Fetches	753,681
Number of Permalinks	3,215,171
Number of Homepages	324,880
Total Compressed Size	25GB
Total Uncompressed Size	148GB
Feeds (Uncompressed)	38.6GB
Permalinks (Uncompressed)	88.8GB
Homepages (Uncompressed)	20.8GB

Table 1: Details of the Blog06 test collection, and its corresponding statistics.

included, covering topics such as news, sports, politics, health, etc. Finally, given the particular severity of spam in the blogosphere, a selection of assumed spam blogs (splogs) were inserted to ensure that Blog track participants had a realistic research setting.

The University of Glasgow monitored the resulting 100,649 blog feeds over an 11 week period from December 2005 to February 2006. During that time, XML feeds, their corresponding homepages and permalink documents were fetched and saved. The final collection was shipped to the Blog track participants by the University of Glasgow[1]. The number of permalinks documents, used as a retrieval unit in the TREC 2006 Blog track, is over 3.2 million of documents. Table 1 shows the statistics of the final collection. Further information about the TREC Blog06 test collection, how it was created, and some of its interesting features compared to other Blog datasets, can be found in [1].

3 Opinion Retrieval Task

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 users information needs seem to be of an opinion, or perspective-finding nature, rather than fact-finding [2]. Incorporating this type of subjectivity in a retrieval context remains a challenge.

3.1 Task

In the TREC 2006 Blog track, the opinion retrieval task involved locating blog posts that express an opinion about a given target. The target can be a "traditional" named entity, e.g. a name of a person, location, or organisation, but also a concept (such as a type of technology), a product name, or an event. The task can be summarised as *What do people think about X*, *X* being a target. The topic of the post was not required to be the same as the target, but an opinion about the target had to be present in the post or one of the comments to the post. For example, for

[1] Further information on obtaining the Blog06 collection can be found at `http://ir.dcs.gla.ac.uk/test_collections/`

the target “skype”, here is an excerpt from a relevant, opinionated blog post:[2]

> *Skype 2.0 eats its young*
> The elaborate press release and WSJ review while impressive don’t help mask the fact that, Skype is short on new ground breaking ideas. Personalization via avatars and ring-tones… big new idea? Not really. Phil Wolff over on Skype Journal puts it nicely when he writes, “If you’ve been using Skype, the Beta version of Skype 2.0 for Windows won’t give you a new Wow! experience.” …

The following is an excerpt from an unopinionated post:[3]

> *Skype Launches Skype 2.0 Features Skype Video*
> Skype released the beta version of Skype 2.0, the newest version of its software that allows anyone with an Internet connection to make free Internet calls. The software is designed for greater ease of use, integrated video calling, and …

While no explicit scenario was associated with the opinion retrieval task, it aims to uncover the public sentiment towards a given entity (the “target”), and hence it can naturally be associated with settings such as tracking consumer-generated content, brand monitoring, and, more generally, media analysis.

3.2 Topics

Topics used in the opinion retrieval task follow the familiar title, description, and narrative structure, as used in topics in other TREC test collections. 50 topics were selected by NIST from a donated collection of queries sent to commercial blog search engines over the time period that the Blogs06 collection was being collected. NIST assessors created the topics by selecting queries, and building topics around those queries. In particular, the *title* fields are the literal queries from the donated search query logs file. Based on the title field, an assessor developed an interpretation of what the searcher who originally submitted the query was looking for. The assessor then searched the Blog06 test collection to see if blog posts with relevant opinions appear in the collection. This searching was by no means complete and no relevance judgements from this phase were retained. Finally, the assessor recorded his/her interpretation of the query in the *description* and *narrative* fields. An example of a topic is included in Figure 1.

3.3 Assessment Procedure

Participants could create queries manually or automatically from the 50 provided topics. They were allowed to submit up to five runs, including a compulsory automatic run using the title-only field of the topic. Moreover, the participants were asked to prioritise runs, in order to define which of their runs would be pooled. Participants were also encouraged to submit manual runs, as such runs are valuable for improving the quality of the test collection. Each submitted run consisted of the top 1,000 opinionated documents for each topic. The *retrieval units* were the documents from the permalinks component of the collection, where there is the post and comments related to it. However, participants were free to use any of the other Blog06 collection components for retrieval such as the XML feeds and/or the HTML homepages.

Pools were formed from the submitted runs of the participants. The two highest priority runs per group were pooled to depth 100. The remaining runs were pooled to depth 10.

[2] Permalink `http://gigaom.com/2005/12/01/skype-20-eats-its-young/`

[3] Permalink `http://www.slashphone.com/115/3152.html`

```
<top>
  <num> Number: 871

  <title> cindy sheehan

  <desc> Description:
  What has been the reaction to Cindy Sheehan and the
  demonstrations she has been involved in?

  <narr> Narrative:
  Any favorable or unfavorable opinions of Cindy Sheehan are
  relevant. Reactions to the anti-war demonstrations she has
  organized or participated in are also relevant.
</top>
```

Figure 1: Blog track 2006, opinion retrieval task, topic 871.

NIST organised the assessments for the opinion retrieval task. However, the relevance judgement of a document for a topic was only made by one assessor, meaning that no assessor disagreement studies could be made. Given a topic and a blog post, assessors were asked to judge the content of the blog post. For the assessment, the *content* of a blog post is defined as the content of the post itself and the contents of all comments to the post. If the relevant content is in a comment, then the permalink is declared to be relevant. Assessments had two levels. The following scale was used for the assessment:

–1 *Not judged.* The content of the post was not examined due to offensive URL or header (such documents do exist in the collection due to spam). Although the content itself was not assessed, it is very likely, given the offensive header, that the post is irrelevant.

0 *Not relevant.* The post and its comments were examined, and does not contain any information about the target, or refers to it only in passing.

1 *Relevant.* The post or its comments contain information about the target, but do not express an opinion towards it. To be assessed as “Relevant”, the information given about the target should be substantial enough to be included in a report compiled about this entity.

If the post or its comments are not only on target, but also contain an explicit expression of opinion or sentiment about the target, showing some personal attitude of the writer(s), then the document had to be judged using the three labels below:

2 Contains an explicit expression of opinion or sentiment about the target, showing some personal attitude of the writer(s), and the opinion expressed is explicitly negative about, or against, the target.

3 Same as (2), but contains both positive and negative opinions.

4 Same as (2), but the opinion expressed is explicitly positive about, or supporting, the target.

Posts that are opinionated, but for which the opinion expressed is ambiguous, mixed, or unclear, were judged simply as “mixed” (3 in the scale).

	Opinion-finding MAP	Topic-relevance MAP
Best	0.3004	0.4219
Median	0.1059	0.1699
Worst	0.0000	2.6e-05

Table 2: Best, median and worst MAP measures for the 57 submitted runs.

Group	Run	MAP	R-prec	bPref	P@10
Univ. of Illinois at Chicago	uicst	**0.1885**	**0.2771**	**0.2693**	**0.5120**
Indiana Univ.	woqs2	0.1872	0.2562	0.2606	0.4340
Tsinghua Univ.	THUBLOGMF	0.1798	0.2647	0.2563	0.3600
Univ. of Amsterdam	UAmsB06All	0.1795	**0.2771**	0.2625	0.4640
CMU (Callan)	blog06r2	0.1576	0.2455	0.2458	0.3580
Univ. of California, Santa Cruz	ucscauto	0.1549	0.2355	0.2264	0.4380
Univ. of Maryland	ParTitDef	0.1547	0.2106	0.2256	0.3360
Univ. of Maryland B.C	UABas11	0.0764	0.1307	0.1202	0.2140
Univ. of Arkansas at Little Rock	UALR06a260r2	0.0715	0.1393	0.1357	0.3320
Univ. of Pisa	pisaBlTit	0.0700	0.1502	0.1535	0.2880
Chinese Academy of Sciences	IIIS	0.0621	0.1134	0.1553	0.2000
National Institute of Informatics	NII1	0.0466	0.1030	0.0851	0.3140
Robert Gordon Univ.	rguOPN	0.0000	0.0004	0.0003	0.0000

Table 3: Opinion retrieval results: the automatic title-only run from each of 13 groups with the best MAP, sorted by MAP. Note that 1 group (Fudan Univ.) did not submit a title-only run. The best in each column is highlighted.

A workable definition of *subjective* or *opinionated* content was proposed. In particular, a post has a subjective content if it contains an *explicit expression of opinion or sentiment about the target, showing a personal attitude of the writer*. Rather than attempting to provide a formal definition, the assessors were given a number of examples, which illustrated the various evaluation labels above.

3.4 Overview of Results

Overall, 14 groups took part in the opinion retrieval task. There were 57 submitted runs, including 53 automatic runs, and 4 manual runs. Each group was asked to submit a compulsory automatic title-only run, for comparison purposes. Of the 57 submitted runs, 27 were pooled to depth 100, and the rest to depth 10.

The metrics used for the opinion retrieval task are mean average precision (MAP), R-Precision (R-Prec), binary Preference (bPref), and Precision at 10 documents (P@10). Since the opinion retrieval task is an adhoc-like retrieval task, the primary measure for evaluating the retrieval performance of the participating groups is the MAP. Table 2 shows the average best, median and worst MAP measures for each topic, across all submitted 57 runs. While these are not “real” runs, they provide a summary of how well the spread of participating systems is performing. Table 3 shows the best-scoring opinion-finding title-only automatic run for each group in terms of MAP, and sorted in decreasing order. R-Prec, bPref and P@10 measures are also shown.

Table 4 shows the best opinion-finding run from each group, in terms of MAP, regardless of the topic length used. Interestingly, none of the manual runs submitted by the participating groups were beneficial to their retrieval performance.

Group	Run	Topics	MAP	R-prec	bPref	P@10
Indiana Univ.	woqln2	TDN	**0.2052**	0.2881	**0.2934**	0.4680
Indiana Univ.	wxoqf2	TDN	0.2019	**0.2934**	0.2824	0.4500
Univ. of Maryland	ParTiDesDmt2	TD	0.1887	0.2421	0.2573	0.3780
Univ. of Illinois at Chicago	uicst	T	0.1885	0.2771	0.2693	**0.5120**
Tsinghua Univ.	THUBLOGMF	T	0.1798	0.2647	0.2563	0.3600
Univ. of Amsterdam	UAmsB06All	T	0.1795	0.2771	0.2625	0.4640
CMU (Callan)	blog06r2	T	0.1576	0.2455	0.2458	0.3580
Univ. of California, Santa Cruz	ucscauto	T	0.1549	0.2355	0.2264	0.4380
Fudan Univ.	mcwil2knl	TDN	0.1179	0.1860	0.1920	0.2940
Univ. of Pisa	pisaBlDes	TD	0.0873	0.1765	0.1620	0.3400
Univ. of Maryland B.C.	UABas11	T	0.0764	0.1307	0.1202	0.2140
Univ. of Arkansas at Little Rock	UALR06a260r2	T	0.0715	0.1393	0.1357	0.3320
Chinese Academy of Sciences	IIIS	T	0.0621	0.1134	0.1553	0.2000
National Institute of Informatics	NII1	T	0.0466	0.1030	0.0851	0.3140
Robert Gordon Univ.	rguOPN	T	0.0000	0.0004	0.0003	0.0000

Table 4: Opinion retrieval results: one run from each of 14 groups with the best MAP, sorted by MAP. Note that all runs in this table were automatic. The best in each column is highlighted. (An extra row was added to show the run (wxoqf2) with the highest R-Prec). T, TD & TDN respectively denote whether the title field, the title and description fields, or the title, description and narrative fields of the topic files were used by the particpant for that run.

In the qrels provided by NIST, documents with adhoc-style relevance to the query (judged 1 or above, as described in Section 3.3 above) were also included. This allows evaluation of the submitted runs based on the relevance of their returned documents. Table 5 reports the best run from each group in terms of topic-relevance. Moreover, Table 6 reports the Spearman's ρ and Kendall's τ correlation coefficients between opinion-finding and topic-relevance measures. The overall rankings of systems on both opinion-finding and topic-relevance measures are extremely similar, as stressed by the obtained high correlations. Figure 2(a) shows a scatter plot of opinion-finding MAP against topic-relevance MAP, which confirms that the correlation is very high.

For the 57 submitted runs, Figure 2(b) plots both opinion-finding MAP and topic-relevance MAP, sorted by opinion-finding MAP. Noticeable from this plot is that runs appear to be clustered into two groups, those above 13% opinion-finding MAP, and those below (see also Table 4). It is interesting that even runs with medium topic-relevance performance can still do comparatively well on opinion-finding MAP compared to runs with stronger topic-relevance performance. In particular, run *uicst* from Univ. of Illinois at Chicago is noticeable as being a strongly performing opinion-finding run compared to its topic-relevance performance.

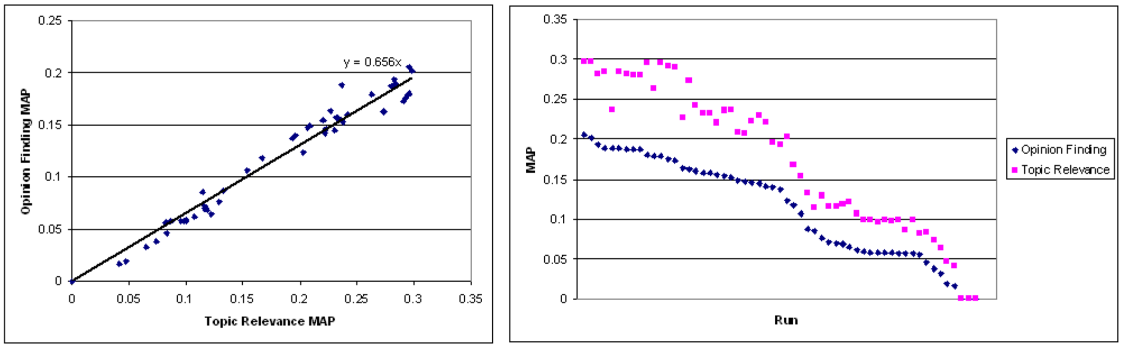
If we rank all the submitted 57 runs by MAP (see Figure 2(b)), for the opinion-finding task, we can determine how many of the top runs are not statistically different, using the Wilcoxon rank test. In particular, of all 57 runs from the opinion-finding MAP, the 9 runs from the best run until run id wxoqs2 (MAP 0.1798) have no significant difference from the best run (woqln2, MAP 0.2052). For topic-relevance MAP, there are some marked differences between the performing systems. In the top 15 runs, 7 are statistically different to the top run, while 8 are not.

Group	Run	Topics	MAP	R-Prec	bPref	P@10
Indiana Univ.	wxoqf2	TDN	**0.2983**	**0.3925**	0.4225	0.6500
Indiana Univ.	woqln2	TDN	0.2963	0.3892	**0.4272**	0.6720
Tsinghua Univ.	THUBLOGMF	T	0.2959	0.3816	0.4177	0.6080
Univ. of Maryland	ParTitDesDef	TD	0.2849	0.3490	0.3998	0.6200
Univ. of Amsterdam	UAmsB06All	T	0.263	0.3674	0.3849	0.6940
Univ. of Illinois at Chicago	uicst	T	0.237	0.3315	0.3415	0.6860
CMU (Callan)	blog06r2	T	0.2324	0.3470	0.3599	0.5480
Univ. of Illinois at Chicago	uicsr	T	0.2267	0.3278	0.3410	**0.7060**
Univ. of California, Santa Cruz	ucscauto	T	0.2203	0.3047	0.3312	0.6480
Fudan Univ.	mcwil2knl	TDN	0.1668	0.2589	0.2826	0.4400
Univ. of Pisa	pisaBlDes	TD	0.1327	0.2329	0.2328	0.5880
Univ. of Maryland B.C.	UABas11	T	0.1288	0.1805	0.1911	0.4520
Univ. of Arkansas at Little Rock	UALR06a500r4	T	0.1192	0.1950	0.1966	0.5180
Chinese Academy of Sciences	IIIS	T	0.1071	0.1903	0.2673	0.3400
National Institute of Informatics	NII1	T	0.0834	0.1522	0.1345	0.5640
Robert Gordon Univ.	rguOPN	T	0.0001	0.0010	0.0010	0.0060

Table 5: Topic-relevance results: documents with 1 or above as relevance label as per the relevance scale defined in Section 3.3. One run from each of 14 groups with the best MAP, sorted by MAP. Note that all runs in this table were automatic. The best in each column is highlighted. (Two extra rows were added to show the runs with the best bPref and P@10, woqln2 and uicsr respectively).

Evaluation Measure	ρ	τ
MAP	0.9745	0.8835
R-Prec	0.9649	0.8609
bPref	0.9505	0.8434
P@10	0.9597	0.8521

Table 6: Correlation of system rankings between opinion-finding performance measures and topic-relevance performance measures. Both Spearman's Correlation Coefficient (ρ) and Kendall's Tau (τ) are reported.



(a) Scatter plot of opinion-finding MAP against topic-relevance MAP.

(b) Opinion finding MAP vs topic-relevance MAP, sorted by opinion-finding MAP.

Figure 2: Figures examining opinion-finding and topic-relevance MAP.

Relevance Scale	Label	Nbr. of Documents	%
Not Judged	-1	0	0%
Not Relevant	0	47491	70.5%
Adhoc-Relevant	1	8361	12.4%
Negative Opinionated	2	3707	5.5%
Mixed Opinionated	3	3664	5.4%
Positive Opinionated	4	4159	6.2%
(Total)	-	67382	100%

Table 7: Relevance assessments of documents in the pool.

Relevance Scale	Nbr. of Splog Documents
Not Judged	0
Not Relevant	8348
Adhoc-Relevant	1004
Negative Opinionated	191
Mixed Opinionated	160
Positive Opinionated	290
(Total)	9993

Table 8: Occurrences of presumed splog documents in the pool

3.5 Overview of the Relevance Judgements

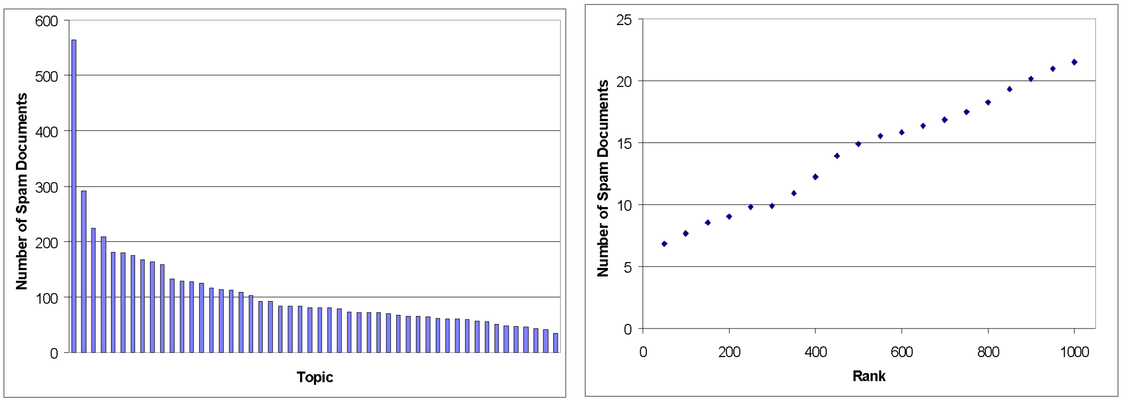
Table 7 shows the breakdown of the relevance assessment of the pooled documents, using the assessment procedure introduced in Section 3.3. About 70% of the pooled documents were judged as irrelevant. As described above, the '-1' element was introduced to allow assessors to discard documents if their associated URL was offensive. However, no assessors made use of this element, choosing in fact to judge all pooled documents. Moreover, it is of note that roughly an equal percentage of opinionated documents were of positive, negative and mixed opinions.

3.5.1 Spam Documents

Since spam is thought to be an issue in the blogosphere, and given that our test collection included a list of assumed splog feeds, we tried to determine the extent to which splog posts had infiltrated the pool, and affected the retrieval systems of the participants. The 17,958 splog feeds in the Blog06 collection generated 509,137 posts. Table 8 provides details on the number of presumed splog posts which infiltrated each element of the relevance scale. In total, 9,993 assumed splog documents were pooled, less than 2% of the splog posts in the collection. Moreover, most assumed splog documents were found not to be opinionated, though those that were were mostly positive.

Figure 3(a) shows the average number of spam documents retrieved by all 57 submitted runs for each topic, in decreasing order. Noticeably, topic 899 (namely "cholesterol") has by far the largest number of splog posts retrieved in the submitted runs (average 564 documents per run). Topic 893 also had a substantial number of splog posts retrieved (average 292 documents per run) - this was again a health topic "zyrtec", which is a medication. Topics which retrieved far fewer spam documents, were concerning people, such as topics 854 and 871 ("Ann Coulter" (34 documents) and "cindy sheehan" (43 documents), respectively).

Next, we examined how the participating systems had been affected by spam documents. Figure 3(b) shows



(a) For each topic, the average number of spam documents retrieved by all of the 57 submitted runs, in decreasing order.

(b) The average number of spam documents retrieved by range of ranks (50), across all topics and submitted runs.

Figure 3: Figures examining the presence of spam documents, by topic and by ranks.

the distribution of spam documents by range of ranks (in units of 50), across all 57 submitted runs and all topics. From this, we can see that on average, systems retrieve more spam documents at later ranks than earlier ranks. In particular, the average number of spam documents retrieved by all systems in the top 10 documents was 1.3. This indicates that the participating systems were good at retrieving non-splog posts at top ranks, and that splog documents were not likely to be retrieved at early ranks. In particular, for the best opinion-finding MAP run of each group, Table 9 shows the mean number of splog documents in the top 10 ranked documents (denoted Spam@10), for all the retrieved documents (Spam@all), and finally BadMAP, which is the Mean Average Precision when the spam documents are treated as the relevant set. BadMAP shows when spam documents are retrieved at early ranks (a low BadMAP value is good, high BadMAP is bad as more spam documents are being retrieved at early ranks). From this table, we can see that some runs were less susceptible to spam documents than others. In particular, runs from the Univ. of Illinois at Chicago and the Univ. of Pisa exhibit the lowest BadMAP values (It is pertinent to note that the Univ. of Pisa reported removing splogs from their collection). In contrast, the run ParTiDesDmt2 of the Univ. of Maryland was affected much more by splog documents.

We also examined the correlation between the official opinion-finding MAP measure calculated using the official relevance assessments (which include spam), and when the assumed spam was removed from the relevance assessments (denoted MAP_NoSpam). Over the 57 submitted runs, the correlation was extremely high ($\rho = 0.9956, \tau = 0.9649$), showing that there is little difference in the overall ranking of submitted runs if the assessors assessed spam or not.

To see if runs that retrieved more spam documents were more likely to be high performing systems or low performing systems, we correlated the ranking of submitted runs by BadMAP, correlating this with MAP_NoSpam. However, the correlation was low ($\rho = 0.2769, \tau = 0.1805$), showing that indeed there was no strong relation between the opinion-finding MAP performance of systems and their likeliness to retrieve spam. However, as the correlation was not negative, it is not the case that low performing systems were more likely to retrieve spam.

Overall, while the Blogs06 test collection contained a component of assumed splogs, the above conclusions suggest that these were not a major hindrance to the retrieval performance of participating groups. Moreover, some topics were more pre-disposed to spam (for example, topics about health), suggesting that these could be identified by statistical predictors.

Group	Run	Spam@10	Spam@all	BadMAP $*10^{-5}$
Indiana Univ.	woqln2	0.78	140.56	6.2
Univ. of Maryland	ParTiDesDmt2	1.92	172.74	14.0
Univ. of Illinois at Chicago	uicst	0.80	38.80	1.0
Tsinghua Univ.	THUBLOGMF	1.56	118.34	5.0
Univ. of Amsterdam	UAmsB06All	0.96	128.22	4.8
CMU (Callan)	blog06r2	0.98	66.02	2.8
Univ. of California, Santa Cruz	ucscauto	1.00	105.38	6.4
Fudan Univ.	mcwil2knl	1.42	60.46	2.4
Univ. of Pisa	pisaBlDes	0.60	48.74	1.6
Univ. of Maryland B.C.	UABas11	1.34	112.74	5.6
Univ. of Arkansas at Little Rock	UALR06a260r2	0.94	95.86	3.0
Chinese Academy of Sciences	IIIS	0.92	75.30	4.2
National Institute of Informatics	NII1	1.44	63.34	4.6
Robert Gordon Univ.	rguOPN	1.56	160.58	6.2

Table 9: Spam measures for runs from Table 4, in the order given. Spam@10 is the mean number of spam posts in the top 10 ranked documents, Spam@all is the mean number of spam posts for each topic. BadMAP is the Mean Average Precision when the spam documents are treated as the relevant set. This shows when spam documents are retrieved at high ranks (low is good, high is bad).

3.5.2 Polarity

We examined the extent to which the submitted runs identified positive and negative opinionated documents. However, because participant systems were not required to rank positively or negatively opinionated documents, the use of precision type measures is not suitable. Therefore, we only look at the recall performance for this analysis. Figure 4 shows the recall of each system in terms of positively opinionated documents against negatively opinionated documents. The gradient of the trend line (0.9342) shows that appears to be a slight overall tendency of the systems to retrieve positively opinionated documents.

Table 10 takes the per-topic best and median runs of the 57 submitted runs, and measures their positive and negative recall. Interestingly, it shows that the best systems are almost equally good at retrieving positive or negative opinions, while the median runs are slightly better at retrieving negatively opinionated documents.

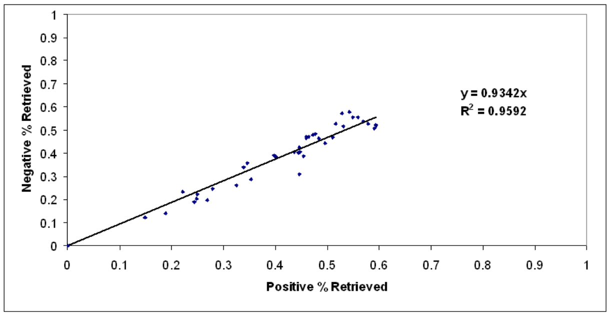


Figure 4: Correlation of positive and negative Recall, by system, over the 57 submitted runs.

	Positive Opinionated Recall	Negative Opinionated Recall
Best Runs	0.7814	0.7754
Median Runs	0.3951	0.4177

Table 10: Recall of positively and negatively opinionated documents, for the per-topic best and median runs.

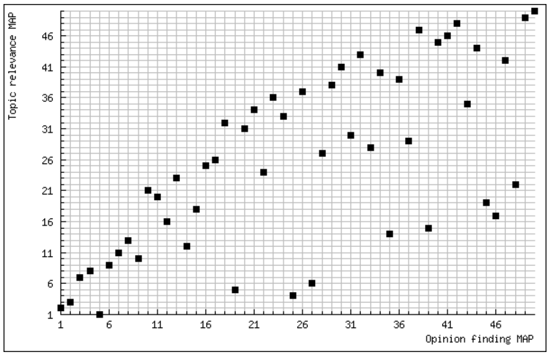


Figure 5: Scatter plot of median opinion-finding MAP against median topical-relevance MAP.

3.5.3 Per Topic Analysis

Analysing further in a per-topic manner allows us to make the following observations. The median number of (topically) relevant documents (i.e., scoring at least 1) per topic was 329, while the median number of documents scoring at least 2 was 182 per topic. The median of the fraction of assessed documents scoring at least 2 vs documents scoring at least 1 was 67% (with minimum 5% and maximum 99%). So overall, for most topics there were more relevant documents than opinionated documents, however, the proportion of opinionated documents varied highly over the topics.

Topics for which median performance (in terms of opinion finding MAP) was high consisted mostly of named entities ("Heineken" (883), "netflix" (863), "Ann Coulter" (854)), whereas low-scoring topics included a mix of such entities along with high-level concepts ("cholesterol" (899), "Business Intelligence Resources" (898)).

As to topics for which the difference between best and median performance was the largest: it seems difficult to define any pattern. These topics are 859, 863, 865, 877, 883, 892 ("letting india into the club," "netflix," "basque," "sonic food industry," "heineken," "jim moran"), and they vary on many dimensions — number of relevant documents, average precision, etc. The same holds true for the topics with the smallest differences between best and median performance: 879, 896, 882, 891, 897 ("hybrid car," "global warming", "seahawks," "intel," "ariel sharon").

Overall, there is a moderate (but not a strong) positive correlation between difficulty in terms of topical relevance and difficulty in terms of opinion finding (Spearman's ρ: 0.6898). Figure 5, which shows a scatter plot of median opinion finding MAP vs median topical MAP (topics sorted by opinion finding MAP), confirms this point. In summary, we infer that the success of the opinion-finding approaches was higher for easier topics.

3.6 Participant Approaches

Looking into the retrieval techniques deployed by the 14 participants, we noticed that most participants approached the opinion retrieval task as a two-stage process. In the first stage, documents are ranked based on topical relevance only, using, mostly, off-the-shelf retrieval systems and weighting models. For example, the University of Maryland, Baltimore County/John Hopkins University (UMBC/JHU) and the Univ. of Arkansas at Little Rock (UALR) used a TF*IDF document ranking scheme. The National Institute of Informatics (NII), the University of Amsterdam (UAmsterdam), the University of Maryland (UMaryland), Chinese Academy of Sciences (CAS), Robert Gordon Univ. and Carnegie Mellon Univ. used language modelling approaches. Finally, the University of Illinois at Chicago (UIllinois) and the University of Pisa (UPisa) used other probabilistic approaches. In the second stage, these results are re-ranked or filtered by applying one or more heuristics for detecting opinions in the documents retrieved at the first stage. The reported approaches by participants for the identification of opinionated content include:

- Dictionary-based approaches: In these approaches, lists of terms and their semantic orientation values were used to rank documents based on the frequency of such words in them, sometimes combined with information about the distance between the sentiment-oriented words and occurrences of query words in the document. In particular, NII proposed a generative language model that models the occurrences of topic terms and opinion-bearing terms in documents. The lists used (e.g., by UMaryland, UIndiana, UAmsterdam, CAS, Tsinghua University (THUIR), UPisa or UMBC/JHU) were either manually-compiled or created automatically. Reports on the success of this approach varied, with some groups observing slight degradation of results compared to their base retrieval scores, and others observing some improvement.

- Text classification approaches: Using training data taken from sources known to contain opinionated content (such as web sites specialising in product reviews) and sources assumed to contain little opinionated content (such as online encyclopedias or news collections), a classifier was trained and used to estimate the degree of opinionated content in retrieved documents. Most groups who used this approach (e.g., UIllinois, UCaliforniaSC, and UALR) favoured Support Vector Machines for their classification, although other classifiers were also used. The success of this approach was limited, possibly because of the difference between training data and the actual opinionated content in blog posts.

- Shallow linguistic approaches: some participants (e.g., UIndiana) used the frequency of pronouns or adjectives as indicators of opinionated content; again, the success of this approach was limited.

In addition to measuring the effect of opinion-detection heuristics, some participants evaluated the benefit of using traditional IR techniques, such as passage retrieval (e.g., UMaryland), or query expansion using pseudo-relevance feedback (e.g., UAmsterdam, UIllinois, or UCaliforniaSC), or using external corpora, (e.g., UIllinois, or UMBC/JHU). Finally, some participants specifically addressed noise in the collection, evaluating the effectiveness of spam detection and other noise removal techniques from the retrieved results (e.g., UIndiana, UPisa, UAmsterdam, or UMBC/JHU). It is difficult to assess the overall effectiveness of these approaches without experimental baselines.

3.7 Baseline Systems

As mentioned above, most participating groups deployed systems using a two-stage process. We desired to assess the usefulness of the post-processing layer in extracting opinionated documents, when compared to a standard IR system. To this end, some additional runs were produced by the organisers, using standard off-the-shelf IR systems, without any opinion-finding specific features. Note that none of these runs were in the assessment pool. Table 11 shows the retrieval performances achieved by the in-house NIST Prise v3, and by the open-source version of Terrier from the University of Glasgow[4]. The indexing settings for Terrier are Porter's stemming and standard

[4] Terrier can be downloaded from `http://ir.dcs.gla.ac.uk/terrier/`

Systems	Topic Fields	Opinion MAP	Relevance MAP
Prise v3	TD	0.1858	0.2908
Terrier v 1.0.2	T	0.1696	0.2703
Terrier v 1.0.2	TD	0.2115	0.3151
Terrier v 1.0.2	TDN	0.1992	0.2892
Terrier v 1.0.2	TN	0.1655	0.2402

Table 11: Performance achieved by standard baseline IR systems. For Terrier, the PL2 weighting model was used with its default parameters.

stopword removal; the DFR PL2 weighting model and its default setting is applied to rank the documents. In particular, it is noticeable that the TD run of Terrier would have achieved the best run on both opinion-finding and topic-relevance MAP measures.

Looking at the 57 submitted runs, there were 42 runs using the title fields of the topics only; 10 using the title and description; and 5 using title, description and narrative. Due to this high variation, it is not possible to draw conclusions as to whether the description and narrative fields helped retrieval for participating systems. However, our baseline runs using Terrier anecdotally suggest that the description field of the topics was beneficial, but the narrative was not.

For future years, participants may benefit from the provision of stronger topic-relevance baseline runs, or detailed instructions on how to use off-the-shelf IR systems, similar to the comparative run systems deployed in the TREC 2006 Terabyte track.

4 Open Task

In the initial proposal for the Blog 2006 track, the intention was to run a time-oriented task, called event timelining, as a second task for TREC 2006. The idea was to focus on the chronological publication order and the associated importance of time in the blogosphere. However, during the TREC 2005 workshop on the Blog track, the workshop participants did not find event timelining to be too interesting, or to be only specific to the blogs. Instead, it was agreed to set up an open task aimed at defining a suitable task for TREC 2007.

Unlike the opinion retrieval task, the open task was not set up as an evaluation task. The open task was meant to provide participants with an opportunity to explore other aspects of blogs besides their opinionated nature. That is, we invited participants to define their own task, which c could be sensibly operationalised and then evaluated in a way that reasonably abstracts the user task. For inspiration, a number of possibilities were sketched in the guidelines for participants, including authority detection (e.g. use part of the corpus to estimate the indegree rank of another part of the corpus), temporal event mining (e.g. identify and follow reactions from bloggers to events which fall under the users' areas of interest), blog finding (e.g. locate blogs about a given topic, rather than posts), spam blog classification, etc.

Participants were asked to propose a "TREC-style" task, which could be used for the TREC 2007 Blog track. This means that the results of the task can be evaluated by a team of assessors and that different approaches can be compared. Groups taking part in the open task were asked to submit a paper describing their proposed task in two steps: First, submit a short abstract, including the definition of the task, some motivation on why it is useful in a realistic blog retrieval environment, and a brief description of the proposed assessment procedure (e.g., how is the task being evaluated?). Secondly, submit a full paper providing a thorough discussion of the proposed task.

4.1 Participants and Results

In total, five proposals were submitted to the open task. They are briefly described below.

NEC Laboratories America proposed the task of identifying spam blogs (splogs) in the collection. Spam blogs are a serious issue in the blogosphere, and their elimination may be a key part in improving results of blog retrieval and other tasks involving blogs. Specifically, the suggested tasks included identification of splogs with fixed training and test sets, and an adaptive splog identification task, where the performance is measured incrementally, as more and more data is available to the system.

The University of Maryland Baltimore County and Johns Hopkins University also proposed the task of splog detection, where the collection is split in time, the first part used for training and the second for testing. Participants would be required to identify splogs in the test collection, and possibly also suggest the type of spamming method being used. An additional extension of the task would evaluate the contribution of spam detection and removal to retrieval performance.

Robert Gordon University proposed a task related to the identification of emerging trends in blogs: topics, which are discussed substantially more during a specific time interval than during preceding intervals. Participants in this task would be given a set of topics and training intervals for each, and would be required to predict those topics that would become "hot topics" during a test interval. Possible approaches to deciding whether a topic is an emerging trend include the volume of discussion about the topic in terms of number of posts or their length, as well as the relation between the topic and other topics.

CSIRO ICT Centre advocated the idea that the availability of more information about a situation and person, i.e. context, will lead to better results for users of search systems. They proposed a sentiment-related task where the blogger's sense-of-self and its changes over time are analysed from the blog posts. In particular, participants in this task would be required to identify, as a first stage, those bloggers who display substantial changes in sense-of-self over time, and, as a second stage, the blog posts which contribute most to tracking these changes. Identifying the blogger's sense-of-self is seen as a partial approach to providing context to the retrieval of blog posts. What this deeper context may add to explicit and implicit search is touched on.

The National Institute of Informatics, Japan proposed a task that is similar to the Story Link Detection at the TDT evaluation, and which involves identifying whether two blog posts discuss the same topic. Participants of this task would be given a set of pairs of blog posts, and would return, for each pair, a decision on whether the two posts are linked — meaning that they share the same topic. Applications of this task include summarisation and "related posts" suggestion.

All the above participants were invited to a (separate) pre-TREC 2006 workshop, to discuss their task proposals. The main purpose of the workshop is to plan the track activities for TREC 2007. While the workshop outcome did not lead to a clear consensus on the submitted proposals, two possible tasks have emerged:

- An information filtering-like task, e.g. *Inform me of new blog entries about X*, *X* being a target as in the opinion retrieval task.

- A Blog expert-like task, e.g. *Find the best blog entries about X, or bloggers with a (recurring) interest in X*, *X* is again a target.

Both above tasks address some interesting features of the blogosphere, and are currently being investigated for a second evaluated task in the TREC 2007 Blog track.

5 Conclusions

TREC 2006 was the first year the Blog track was run. A new large test collection of blog data, called Blog06, was created, and a particular feature of blogs has been tackled, namely the opinionated nature of posts on the blogosphere. The participants results suggest that this task is challenging, and requires further investigation. We found that the retrieval performance on the opinion retrieval task is strongly dominated by the performance on the underlying topic relevance task, emphasising the importance of a strong retrieval baseline. We also found that the pooled documents were not infiltrated by spam to any great extent, and the presence of the spam in the pool did not affect the overall ranking of systems. Moreover, there was no strong evidence that the participating systems retrieved one kind of opinion over another. Finally, there seems to be a positive but not strong correlation between difficulty of topics, in terms of opinion-finding MAP and topic-relevance MAP. It is hoped that by using the relevance assessments of this year as training data, participants will be able to further their techniques for identifying opinionated blog posts.

For the open task, there was no clear emerging task suitable for evaluation in TREC 2007. However, an information filtering or blog identification task (as discussed in Section 4.1) seem to address some interesting elements of the blogosphere, and could be run in forthcoming Blog tracks.

Task details for TREC 2006 Blog track are maintained on the track wiki, at `http://www.science.uva.nl/research/iiwiki/wiki/index.php/TREC-blog`.

Details on the TREC 2007 Blog track are provided on the following wiki page: `http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG`

Acknowledgements

We are grateful to Nielsen BuzzMetrics/BlogPulse for contributing a feed list for creating the Blog06 test collection. We also would like to thank Matthew Hurst and Janyce Wiebe for useful advice and discussions on setting up the opinion retrieval task. We would like to thank Ali Azimi Bolourian for monitoring some of the crawls during the Blog06 test collection creation. Finally, we would like to thank Gianni Amati for various comments on the Blog track.

Gilad Mishne and Maarten de Rijke were supported by the Netherlands Organization for Scientific Research (NWO) under project number 220-80-001.

References

[1] Craig Macdonald and Iadh Ounis. The TREC Blog06 Collection : Creating and Analysing a Blog Test Collection *DCS Technical Report TR-2006-224*. Department of Computing Science, University of Glasgow. 2006. `http://www.dcs.gla.ac.uk/~craigm/publications/macdonald06creating.pdf`

[2] Gilad Mishne and Maarten de Rijke. A Study of Blog Search. In *Proceedings of ECIR-2006*. LNCS vol 3936. Springer 2006.

[3] Amanda Lenhart and Susannah Fox. Bloggers : a portrait of the Internet's new storytellers *Pew Internet & American Life Project*. July. 2006.