Overview of the TREC-2007 Blog Track

Craig Macdonald, Iadh Ounis
University of Glasgow
Glasgow, UK
{craigm,ounis}@dcs.gla.ac.uk

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

1. INTRODUCTION

The goal of the Blog track is to explore the information seeking behaviour in the blogosphere. It aims to create the required infrastructure to facilitate research into the blogosphere and to study retrieval from blogs and other related applied tasks. The track was introduced in 2006 with a main opinion finding task and an open task, which allowed participants the opportunity to influence the determination of a suitable second task for 2007 on other aspects of blogs besides their opinionated nature. As a result, we have created the first blog test collection, namely the TREC Blog06 collection, for adhoc retrieval and opinion finding. Further background information on the Blog track can be found in the 2006 track overview [2].

TREC 2007 has continued using the Blog06 collection, and saw the addition of a new main task and a new subtask, namely a blog distillation (feed search) task and an opinion polarity subtask respectively, along with a second year of the opinion finding task. NIST developed the topics and relevance judgments for the opinion finding task, and its polarity subtask. For the blog distillation task, the participating groups created the topics and the associated relevance judgments. This second year of the track has seen an increased participation compared to 2006, with 20 groups submitting runs to the opinion finding task, 11 groups submitting runs to the polarity subtask, and 9 groups submitting runs to the blog distillation task. This paper provides an overview of each task, summarises the obtained results and draws conclusions for the future.

The remainder of this paper is structured as follows. Section 2 provides a short description of the used Blog06 collection. Section 3 describes the opinion finding task and its polarity subtask, providing an overview of the submitted runs, as well as a summary of the main used techniques by the participants. Section 4 describes the newly created blog distillation (feed search) task, and summarises the results of the runs and the main approaches deployed by the participating groups. We provide concluding remarks in Section 5.

2. THE BLOG06 TEST COLLECTION

All tasks in the TREC 2007 Blog track use the Blog06 collection, representing a large sample crawled from the blogosphere over an eleven week period from December 6, 2005 until February 21, 2006. The collection is 148GB in size, with three main components consisting of 38.6GB of XML feeds (i.e. the blog), 88.8GB of permalink documents (i.e. a single blog post and all its associated comments) and 28.8GB of HTML homepages (i.e. the main entry to the blog). In order to ensure that the Blog track experiments are conducted in a realistic and representative setting, the collection also includes spam, non-English documents, and some non-blogs documents such as RSS feeds.

The number of permalink documents in the collection is over 3.2 million, while the number of feeds is over 100,000 blogs. The permalink documents are used as a retrieval unit for the opinion finding task and its associated polarity subtask. For the blog distillation task, the feed documents are used as the retrieval unit. The collection has been distributed by the University of Glasgow since March 2006. Further information on the collection and how it was created can be found in [1].

3. OPINION FINDING TASK

Many blogs are created by their authors 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. The opinion finding task is an articulation of a user search task, where the information need seems to be of an opinion, or perspective-finding nature, rather than fact-finding. 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. This is the second running of the opinion finding task in the Blog track. This year, it was the most popular task of the track, with 20 participating groups.

3.1 Topics and Relevance Judgments

Similar to TREC 2006, the opinion retrieval task involved locating blog posts that express an opinion about a given target [2]. 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 is 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, as identified by the permalink.

Topics used in the opinion finding task follow the familiar title, description, and narrative structure, as used in topics in other TREC test collections. 50 topics were again selected by NIST from a larger query log obtained from a commercial blog search engine. The topics were created by NIST using the same methodology as last year, namely selecting queries from the query log, and building topics around those queries [2]. An example of a TREC 2007 topic is included in Figure 1.

3.2 Pooling and Assessment Procedure

Participants could create queries manually or automatically from the 50 provided topics. They were allowed to submit up to six runs, including a compulsory automatic run using the title field of

```
<top>
  <num> Number: 930 </num>

  <title> ikea  </title>

  <desc> Description:
 Find opinions on Ikea or its products.
</desc>
  <narr> Narrative:
  Recommendations to shop at Ikea are
relevant opinions. Recommendations of
Ikea products are relevant opinions.
Pictures on an Ikea-related site that
are not related to the store or its
products are not relevant.
</narr>
</top>
```

Figure 1: Blog track 2007, opinion retrieval task, topic 930.

the topics, and another compulsory automatic run, using the title field of the topics, but with all opinion-finding features turned off. The latter was required to draw further conclusions on the extent to which a strong topic relevance baseline is required for an effective opinion retrieval system. It also helps to draw conclusions on the real effectiveness of the specifically used opinion finding approaches.

As mentioned in Section 2, for the purposes of the opinion finding task, the document retrieval unit in the collection is a single blog post plus all of its associated comments as identified by a permalink. 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.

Overall, 20 groups participated in the opinion finding task, submitting 104 runs, including 98 automatic runs and 6 manual runs. The participants were asked to prioritise runs, in order to define which of their runs would be pooled. Like in TREC 2006, the guidelines of the Blog track encouraged participants to submit manual runs to improve the quality of the test collection. Each submitted run consisted of the top 1,000 opinionated documents (permalinks) for each topic. NIST formed the pools from the submitted runs using the three highest-priority runs per group, pooled to depth 80. In case of ties, the manual runs were preferred over the automatic runs, and among the automatic title-only tied runs, the compulsory ones were preferred.

NIST organised the relevance assessments for the opinion finding task, using the same assessment procedure defined in 2006 [2], with some further tightening up of the guidelines given to the assessors. In particular, the assessment procedure had two levels. The first level assesses whether a given blog post, i.e. a permalink, contains information about the target and is therefore relevant. The second level assesses the opinionated nature of the blog post, if it was deemed relevant in the first assessment level. Given a topic and a blog post, assessors were asked to judge the content of the blog posts. The following scale was used for the assessment:

0 *Not relevant.* The post and its comments were examined, and do 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.

Relevance Scale	Label	Nbr. of Documents	%
Not Relevant	0	42434	77.7%
Adhoc-Relevant	1	5187	9.5%
Negative Opinionated	2	1844	3.4%
Mixed Opinionated	3	2196	4.0%
Positive Opinionated	4	2960	5.4%
(Total)	-	54621	100%

Table 1: Relevance assessments of documents in the pool.

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

2 *Negatively opinionated.* 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 *Mixed.* Same as (2), but contains both positive and negative opinions.

4 *Positively opinionated.* 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).

Table 1 shows a breakdown of the relevance assessment of the pooled documents, using the assessment procedure described above. About 78% of the pooled documents were judged as irrelevant. Moreover, there were roughly an equal percentage of negative and mixed opinionated documents, but slightly more positive opinionated documents, suggesting that overall, the bloggers had more positive opinions about the topics tackled by the TREC 2007 opinion finding topics set. Figure 2 shows the number of relevant positive and negative opinionated documents for each topic. Topic "northernvoice" (914) or topic "mashup camp" (925) have only relevant positive opinionated documents in the pool, whereas topic "censure" (943) or topic "challenger" (923) have more negative than positive opinionated documents in the pool, perhaps illustrating the nature of these tackled topics.

3.3 Overview of Results

Since the opinion finding task is an adhoc-like retrieval task, the primary measure for evaluating the retrieval performance of the participating groups is the mean average precision (MAP). Other metrics used for the opinion finding task are R-Precision (R-Prec), binary Preference (bPref), and Precision at 10 documents (P@10).

Table 2 provides the average best, median and worst MAP measures for each topic, across all submitted 104 runs. While these are not "real" runs, they provide a summary of how well the spread of participating systems is performing. In particular, it is of interest to note that the retrieval performances of the participating groups in TREC 2007 are markedly higher than those reported in TREC 2006 on the same task. For example, the median MAP measure of the submitted runs for the opinion finding task has increased from 0.1059 in TREC 2006 [2] to 0.2416 in TREC 2007. Further investigation is required in order to conclude whether this is due to the TREC 2007 topics being easier than those used in TREC 2006, or if

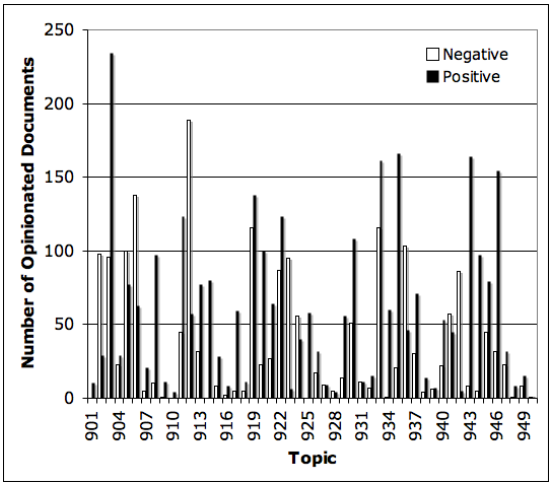


Figure 2: Number of positive and negative opinionated documents per topic in the pool.

	Opinion-finding MAP	Topic-relevance MAP
Best	0.5182	0.6382
Median	0.2416	0.3340
Worst	4.2e-05	0.0001

Table 2: Best, median and worst MAP measures for the 104 submitted runs to the opinion finding task.

the increase is due to the use of more effective retrieval approaches by the participants.

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 reported. Table 4 shows the best opinion-finding run from each group, in terms of MAP, regardless of the topic length used.

Each participating group was required to submit a compulsory automatic run, using only the title field of the topics, with all opinion finding features of the retrieval system turned off (i.e. a *topic-relevance baseline* run). The idea is to have a better understanding of the actual effectiveness of the opinion detection approaches deployed by the participating groups, allowing to draw conclusions as to whether the used opinion finding techniques actually help retrieving opinionated documents. Table 5 shows the best baseline run from each group, in terms of opinion-finding MAP. Comparing Tables 3 and 5, it is interesting to note that only one of the top five performing opinion finding runs was actually a topic-relevance baseline run. In particular, out of the 5 best opinion-finding performing runs in Table 3, only run uams07topic from the University of Amsterdam was a topic-relevance run.

In order to assess which opinion finding features and approaches deployed by the participating groups have actually worked, we compare the performance of the best performing opinion finding run of each group to its best submitted topic-relevance baseline. A relative increase in performance indicates that the used opinion finding features were useful. A relative decrease in performance indicates that the deployed opinion finding features did not help in retrieval. Table 6 shows the improvements of the best submitted compulsory automatic title-only runs over the baselines. Note that the best performing group on the opinion finding task, namely the UIC group, did not officially submit a baseline run, making it difficult to conclude on the success of their deployed opinion finding features. It

Evaluation Measure	ρ	τ
MAP	0.9778	0.8813
R-Prec	0.9677	0.8518
bPref	0.8118	0.9448
P@10	0.8032	0.9366

Table 8: 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.

is interesting to note that the best opinion finding run by the University of Amsterdam has decreased the performance of the their strongly performing uams07topic topic-relevance baseline by over 57%. On the other hand, the opinion finding features used by the University of Glasgow, Indiana University, and the University of Arkansas at Little Rock seem to be helpful, improving their performance on the task by 15.8%, 14% and 13.9%, respectively, depsite their good performing baselines.

Given the two levels assessment procedure, it is possible to evaluate the submitted runs in a classical adhoc fashion, i.e. based on the relevance of their returned documents (judged 1 or above, as described in Section 3.2 above). Table 7 reports the best run from each group in terms of topic-relevance, regardless of the topic length.

Moreover, Table 8 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 very similar, as stressed by the obtained high correlations. A similar finding was observed in TREC 2006 [2], suggesting again that good performances on the opinion finding task are strongly dominated by good performances on the underlying topic-relevance task. Figure 3(a) shows a scatter plot of opinion-finding MAP against topic-relevance MAP, which confirms that the correlation is very high.

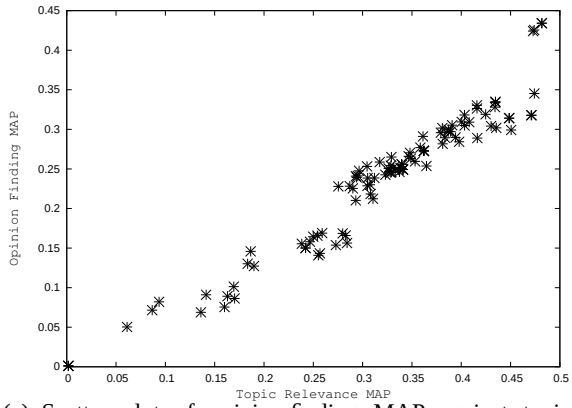
Finally, we report on the extent to which the 17,958 presumed splog feeds and their associated 509,137 spam posts, which were injected into the Blog06 collection during its creation have infiltrated the pool. Table 9 provides details on the number of presumed splog posts which infiltrated each element of the relevance scale. In total, 7,086 assumed splog documents were pooled, less than 1.5% of the splog posts in the collection. Moreover, there was a roughly equal number of relevant only and opinionated splog posts, though those that were opinionated were mostly positive. Figure 4 shows the average number of spam documents retrieved by all 104 submitted runs for each topic, in decreasing order.

Noticeably, unlike in last year's TREC 2006 topics set where the most spammed topics where about health, we note that topic 915 (namely "allianz") had by far the largest number of splog posts retrieved in the submitted runs (average 703 documents per run). Topic "grammys" (936) and topic "teri hatcher" also had a substantial number of splog posts retrieved (average 466 and 309 documents per run, respectively). These are widely popular topics, which might be prone to being spammed. Similar to TREC 2006 though, topics which retrieved far fewer spam documents, were concerning people not featuring in the tabloid news as often, such as topics 924 and 904: "mark driscoll" (23 documents) and "alterman" (9 documents), respectively.

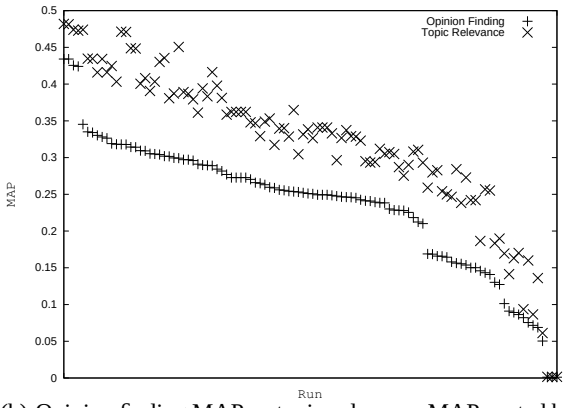
Next, we examined how the participating systems had been affected by spam documents. Table 10 shows the mean number of splog documents in the top 10 ranked documents (denoted Spam@10), and for all the retrieved documents (Spam@all). The table also reports BadMAP, which is the Mean Average Precision when the pre-

Group	Run	MAP	R-prec	b-Bref	P@10
UIC (Zhang)	uic1c	**0.4341**	**0.4529**	**0.4724**	**0.690**
UAmsterdam (deRijke)	uams07topic	0.3453	0.3872	0.3953	0.562
UGlasgow (Ounis)	uogBOPFProxW	0.3264	0.3657	0.3497	0.552
DalianU (Yang)	DUTRun2	0.3190	0.3671	0.3686	0.600
FudanU (Wu)	FDUTOSVMSem	0.3143	0.3465	0.3499	0.460
CAS (Liu)	Relevant	0.3041	0.3600	0.3779	0.446
UArkansas Littlerock (Bayrak)	UALR07BlogIU	0.2911	0.3263	0.3134	0.580
IndianaU (Yang)	oqsnr2opt	0.2894	0.3572	0.3419	0.532
UNeuchatel (Savoy)	UniNEblog1	0.2770	0.3353	0.3074	0.492
FIU (Netlab team)	FIUbPL2	0.2728	0.3204	0.2925	0.454
UWaterloo (Olga)	UWopinion3	0.2631	0.3344	0.2980	0.496
Zhejiangu (Qiu)	EAGLE1	0.2561	0.3159	0.2867	0.428
CAS (NLPR-IACAS)	NLPRPST	0.2542	0.3168	0.2945	0.462
BUPT (Weiran)	prisOpnBasic	0.2466	0.3018	0.2835	0.456
KobeU (Eguchi)	KobePrMIR01	0.246	0.3011	0.2744	0.440
NTU (Chen)	NTUAutoOp	0.2282	0.2614	0.2577	0.464
KobeU (Seki)	Ku	0.1689	0.2417	0.2190	0.254
RGU (Mukras)	rgu0	0.1686	0.2266	0.2163	0.288
UBuffalo (Ruiz)	UB2	0.1013	0.1297	0.1238	0.144
Wuhan (Lu)	NOOPWHU1	0.0011	0.0071	0.0072	0.008

Table 3: Opinion finding results: the automatic title-only run from each of 20 groups with the best MAP, sorted by MAP. The best in each column is highlighted.



(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 3: Figures examining opinion-finding and topic-relevance MAP.

Relevance Scale	Nbr. of Splog Documents
Not Relevant	6357
Adhoc-Relevant	361
Negative Opinionated	78
Mixed Opinionated	98
Positive Opinionated	192
(Total)	7086

Table 9: Occurrences of presumed splog documents in the opinion finding task pool.

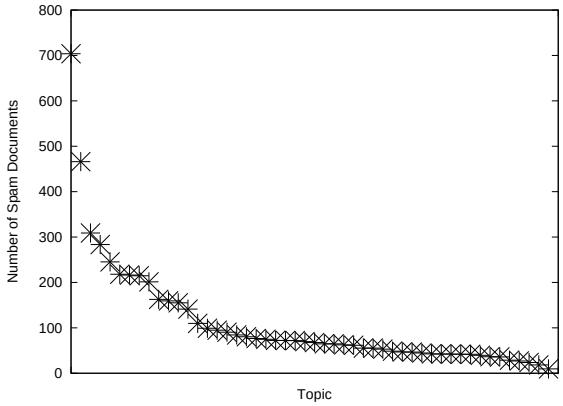


Figure 4: Distribution of number of spam documents retrieved per topic.

Group	Run	Automatic	Fields	MAP	R-prec	b-Bref	P@10
UIC (Zhang)	uic1c	yes	T	**0.4341**	**0.4529**	**0.4724**	**0.690**
UAmsterdam (deRijke)	uams07topic	yes	T	0.3453	0.3872	0.3953	0.562
IndianaU (Yang)	oqlr2fopt	yes	TDN	0.3350	0.3925	0.378	0.576
UGlasgow (Ounis)	uogBOPFProxW	yes	T	0.3264	0.3657	0.3497	0.552
DalianU (Yang)	DUTRun2	yes	T	0.3190	0.3671	0.3686	0.600
FudanU (Wu)	FDUTisdOpSVM	yes	T	0.3179	0.3467	0.3501	0.454
FIU (Netlab team)	FIUDDPH	yes	TD	0.3053	0.3498	0.3475	0.492
UNeuchatel (Savoy)	UniNEblog3	yes	TD	0.3049	0.3438	0.3266	0.516
CAS (Liu)	Relevant	yes	T	0.3041	0.3600	0.3779	0.446
UArkansas Littlerock (Bayrak)	UALR07BlogIU	yes	T	0.2911	0.3263	0.3134	0.580
UWaterloo (Olga)	UWopinion3	yes	T	0.2631	0.3344	0.298	0.496
CAS (NLPR-IACAS)	NLPRPTD2	yes	TD	0.2587	0.3088	0.2956	0.456
Zhejiangu (Qiu)	EAGLE1	yes	T	0.2561	0.3159	0.2867	0.428
BUPT (Weiran)	prisOpnBasic	yes	T	0.2466	0.3018	0.2835	0.456
KobeU (Eguchi)	KobePrMIR01	yes	T	0.2460	0.3011	0.2744	0.440
NTU (Chen)	NTUManualOp	no	T	0.2393	0.2659	0.2749	0.486
KobeU (Seki)	Ku	yes	T	0.1689	0.2417	0.219	0.254
RGU (Mukras)	rgu0	yes	T	0.1686	0.2266	0.2163	0.288
UBuffalo (Ruiz)	UB1	yes	TDN	0.1501	0.2001	0.1887	0.266
Wuhan (Lu)	NOOPWHU1	yes	T	0.0011	0.0071	0.0072	0.008

Table 4: Opinion finding results: best run from each of the 20 groups, regardless of the used topic length. The best in each column is highlighted.

sumed spam documents are treated as the relevant set. BadMAP shows when spam documents are retrieved at early ranks (a low BadMAP value is good, while a 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, the run from UIC exhibited a perfect 0 BadMAP and the lowest Spam@10 and Spam@all measures, suggesting that this group has very successfully applied splog detection techniques (Indeed, UIC has experimented with a spam detection module in TREC 2007). In contrast, the run NTUAutoOp from NTU was affected much more by splog documents.

To see if runs that retrieved less 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 opinion finding MAP. However, the correlation was low ($\rho = 0.01, \tau = 0.03$), showing that for this task, systems which did remove spam documents were not any more likely to have a higher opinion retrieval performance.

3.4 Polarity Subtask

The polarity subtask was introduced in TREC 2007 as a natural extension of the opinion task, and was intended to represent a text classification-related task, requiring participants to determine the polarity (or orientation) of the opinions in the retrieved documents, namely whether the opinions in a given document are positive, negative or mixed. Participants were encouraged to use last years 50 opinion task queries, with their associated relevance judgments for training. Indeed, during the assessment procedure in the TREC 2006 blog track, for each document in the pool, the NIST assessors have specified the polarity of the relevant documents as described in Section 3.2 above: relevant negative opinion (judged as 2 in qrels); relevant mixed positive and negative (judged as 3 in qrels); relevant positive opinion (judged as 4 in qrels).

Groups participating in the opinion task and wishing to submit runs to the polarity subtask were asked to provide a corresponding and separate file for a submitted run to the opinion task, which details the predicted polarity for each retrieved document for each query. Submitted runs included the same documents in the same order as for the opinion finding runs, but with an additional polarity predictive label. Overall, 11 groups submitted 38 runs to the polarity subtask, including 32 automatic runs and 6 manual runs.

	R-Acc
Best	0.2959
Median	0.1227
Worst	0.0004

Table 11: Best, median and worst R-accuracy measures for the 38 submitted runs to the polarity subtask.

The initial intention was to evaluate the submitted runs using a classification accuracy measure (i.e. set precision). However, a measure like classification accuracy is comparable between runs only when every run classifies every document in the test set. In the polarity subtask, each run only provides a classification for the documents in its associated ranked opinion finding run. This presents three problems: not every run classifies the same documents, the treatment of unclassified documents is undefined, and no standard cutoff in the ranking is apparent.

To provide scores that are suitably comparable between runs, we report a measure called "R-accuracy" (R-Acc). This is the fraction of retrieved documents above rank R that are classified correctly, where R is the number of opinion-containing documents for that topic. The proposed measure is analogous to R-precision where only the correctly-classified opinion documents are counted as relevant. We also report accuracy at fixed rank cutoffs (A@10 and A@1000) as a secondary metric. For all measures, unjudged retrieved documents have no correct classification. The assumption is that if a submitted run had known that the document was not opinionated then the run should not have retrieved it, i.e. by retrieving it the run assumes that the document was opinionated, and hence must have wrongly classified it. Table 11 provides the average best, median and worst R-Acc measures for each topic, across all submitted 38 runs.

Group	Run	MAP	R-prec	b-Bref	P@10
UAmsterdam (deRijke)	uams07topic	**0.3453**	**0.3872**	**0.3953**	**0.562**
FudanU (Wu)	FDUNOpRSVMT	0.3178	0.3447	0.3498	0.452
CAS (Liu)	Relevant	0.3041	0.3600	0.3779	0.446
DalianU (Yang)	DUTRun1	0.2890	0.3368	0.3249	0.502
UGlasgow (Ounis)	uogBOPFProx	0.2817	0.3366	0.3098	0.454
UNeuchatel (Savoy)	UniNEblog1	0.277	0.3353	0.3074	0.492
FIU (Netlab team)	FIUbPL2	0.2728	0.3204	0.2925	0.454
Zhejiangu (Qiu)	EAGLE1	0.2561	0.3159	0.2867	0.428
UArkansas Littlerock (Bayrak)	UALR07Base	0.2554	0.3145	0.2867	0.440
IndianaU (Yang)	oqsnr1Base	0.2537	0.323	0.3091	0.446
CAS (NLPR-IACAS)	NLPRPTONLY	0.2506	0.3166	0.2917	0.452
UWaterloo (Olga)	UWbasePhrase	0.2486	0.3087	0.2861	0.432
BUPT (Weiran)	prisOpnBasic	0.2466	0.3018	0.2835	0.456
NTU (Chen)	NTUAuto	0.2254	0.2795	0.2588	0.412
KobeU (Seki)	Ku	0.1689	0.2417	0.219	0.254
RGU (Mukras)	rgu0	0.1686	0.2266	0.2163	0.288
Wuhan (Lu)	NOOPWHU1	0.0011	0.0071	0.0072	0.008

Table 5: Opinion finding results: automatic title-only baseline runs from each of the group with the best MAP, sorted by MAP. In these runs, all opinion finding features are switched off. The best in each column is highlighted. Note that some groups did not submit the compulsory automatic title-only baseline run.

Table 12 shows the best-scoring title-only polarity detection run for each group in terms of R-accuracy, and sorted in decreasing order of R-accuracy, while Table 13 shows the same information, but regardless of the topic length. Noticeable from these tables is that the runs appear to be clustered into two groups, those above 11% polarity detection R-accuracy, and those below.

It is interesting to note that the Spearman's ρ and Kendall's τ correlation coefficients between the polarity detection R-accuracy results and their corresponding opinion-finding MAP results over the 38 submitted polarity runs are very high (ρ = 0.9345 and τ =0.8065). This can be explained by the fact that the systems which are more successful at retrieving opinionated documents ahead of relevant ones, will then have more documents for which they can make a correct classification. Systems which perform poorly at retrieving opinionated documents are by definition not going to have the chance to classify as many documents correctly, hence the strong correlation is expected.

3.5 Participant Approaches

There were a wide range of deployed techniques by the participating groups. In this section, we focus on those groups whose use of opinion finding features have markedly improved their topic-relevance baseline as shown in Table 6. Looking into the main features of the best submitted runs, we note the following:

Indexing All the participating groups only indexed the Permalink component of the Blog06 collection, but the group from the University of Waterloo, which used all three components of the collection namely, Permalinks, Feeds and Homepages.

Retrieval Similar to TREC 2006, most of the participating groups used a two-stage approach for document retrieval [2]. In the first stage, documents are ranked using a variety of document weighting models ranging from BM25 (e.g. University of Indiana and University of Waterloo) to Divergence From Randomness models (e.g. University of Glasgow and FIU (Netlab team)), through language modelling (e.g. University of Amsterdam). Many participants used off-the-shelf systems such as Indri or Terrier. In the second stage of the retrieval process, the retrieved documents are re-ranked taking into account opinion finding features, often through a combination of scores mechanism.

Opinion Finding Features From looking at the results, we observe that there were two main effective approaches for detecting opinionated documents, which both led to improvements over a topic-relevance baseline. The first approach, used for example by the University of Glasgow and FIU, consists in automatically building a weighted dictionary from the relevance assessments of the TREC 2006's opinion finding task. The weight of each term in the dictionary estimates its opinionated discriminability. The weighted dictionary is then submitted as a query to generate an opinionated score for each document of the collection. The second approach, tested for example by the University of Arkansas at Little Rock and the University of Waterloo, uses a pre-compiled list of subjective terms and indicators and re-ranks the documents based on the proximity of the query terms to the aforementioned pre-compiled list of terms.

In the following, we provide more details on methods used by the 5 best performing groups, whose approaches for detecting opinionated documents have worked well, compared to a topic-relevance baseline as shown in Table 6:

The **University of Glasgow (UoG)** experimented with two approaches for detecting opinionated documents, integrated into their Terrier search engine. The first purely statistical approach uses a compiled English word list collected from various available linguistic resources. UoG measured the opinionated discriminability of each term in the word list using an information theoretic divergence measure based on the relevance assessments of the TREC 2006's opinion finding task. They have then estimated the opinionated nature of each document in the collection with the PL2 Divergence from Randomness (DFR) weighting model, and using the weighted opinionated word list as a query. The same approach was used to detect polarity. Their second opinion detection approach uses OpinionFinder, a freely available toolkit, which identifies subjective sentences in text. For a given document, they adapted OpinionFinder to produce an opinion score for each document, based on the identified opinionated sentences. Using either of two opinion

Group	Best Baseline	Baseline MAP	Best Non-baseline	Non Baseline MAP	% Increase
UGlasgow (Ounis)	uogBOPFProx	0.2817	uogBOPFProxW	**0.3264**	**15.87%**
IndianaU (Yang)	oqsnr1Base	0.2537	oqsnr2opt	0.2894	14.07%
UArkansas Littlerock (Bayrak)	UALR07Base	0.2554	UALR07BlogIU	0.2911	13.98%
DalianU (Yang)	DUTRun1	0.289	DUTRun2	0.319	10.38%
UWaterloo (Olga)	UWbasePhrase	0.2486	UWopinion3	0.2631	5.83%
CAS (NLPR-IACAS)	NLPRPTONLY	0.2506	NLPRPST	0.2542	1.44%
NTU (Chen)	NTUAuto	0.2254	NTUAutoOp	0.2282	1.24%
FudanU (Wu)	FDUNOpRSVMT	0.3178	FDUTisdOpSVM	0.3179	0.03%
FIU (Netlab team)	FIUbPL2	0.2728	FIUdPL2	0.2728	0.00%
Wuhan (Lu)	NOOPWHU1	0.0011	OTWHU101	0.0011	0.00%
KobeU (Seki)	Ku	0.1689	KuKnn	0.1657	-1.89%
Zhejiangu (Qiu)	EAGLE1	0.2561	EAGLE2	0.2493	-2.66%
CAS (Liu)	Relevant	0.3041	DrapOpi	0.1659	-45.45%
RGU (Mukras)	rgu0	0.1686	rgu2	0.0892	-47.09%
UAmsterdam (deRijke)	uams07topic	**0.3453**	uams07mmqop	0.1459	-57.75%
BUPT (Weiran)	prisOpnBasic	0.2466	prisOpnC2	0.0821	-66.71%

Table 6: What worked. Improvements over the baselines, for automatic title-only runs. The best in each column is highlighted. Some groups did not submit title-only baseline runs (e.g. UIC group), and some did not submit any run with specific opinion finding features (e.g. UNeuchatel).

detection approaches, UoG used the opinionated scores of the documents as prior evidence, and integrated them with the relevance scores produced by the document weighting model used. All their six submitted runs used the PL2F field-based weighting model. One of their topic-relevance baselines included a DFR-based proximity model. They found that the use of the word list-based statistical opinion detection approach markedly improved their topic-relevance only baseline, leading to a substantial and marked improvement of 15.8% compared to the topic-relevance baseline (run uogBOPFProxW vs run uogBOPFProx). Interestingly, they also found that the opinion finding technique based on the OpinionFinder tool was as effective as the statistical word list-based approach, although it was less efficient. They also reported that the use of proximity search is helpful.

The **University of Indiana (IndianaU)** focused on combining multiple sources of evidence to detect opinionated blog postings. Their approach to opinion blog retrieval consisted of first applying traditional retrieval methods to retrieve on-topic blogs and then boosting the ranks of opinionated blogs based on combined opinion scores generated by multiple assessment methods. Indiana's opinion assessment/detection method is comprised of High Frequency Module, which identifies opinion blogs based on the frequently used opinion terms, low frequency module, which leverages uncommon/rare term patterns (e.g., 'sooo good') for expressing opinions, IU Module, which makes use of 'I' and 'You' collocations (e.g. 'I believe') that qualify opinion sentences, Wilson's lexicon module, which makes use of Wilson's subjective lexicons, and opinion acronym module, which utilises the small set of opinion acronyms (e.g., 'imho') that are likely to be missed by preceding modules. Indiana's training data consisted of TREC 2006's opinion finding relevance data supplemented by the external IMDB movie review data, both of which were used to tune their opinion scoring and fusion module in an interactive system optimisation mechanism called the Dynamic Tuning Interface. All of the lexicon terms were scored with positive and negative values, which facilitated their participation in the polarity subtask. They found that their opinion finding approach improves upon the topic-relevance only baseline.

The **University of Arkansas at Little Rock (UArkansas)** used various opinion finding heuristics on top of a topic-relevance baseline. Their best performing opinion finding run re-ranked the documents returned by the baseline, by taking into account the proximity of words such as "I", "you", "me", "us", we" and opinion indicator words such as "like", "feel","think","hate" to the actual query words. They found that such a simple proximity-based approach could markedly improve the opinion finding retrieval effectiveness of their topic relevance baseline (about 14% improvement). UArkansas also experimented with a machine learning-based approach, which re-ranks the baseline results by associating a category to the queries. This approach while slightly improving upon the performance of the topic-relevance baseline, was comparatively less successful than the proximity-based approach.

The **Dalian University of Technology (DUT)** filtered out all non-English blog posts during indexing. They used an external resource, namely the Wikipedia, and a manually built sentiment lexicon resource to find opinions. In the polarity subtask, DUT used a method based on SVM, to assess the polarity of the retrieved blog posts. Judging by the results, DUT found that their used sentiment resources had improved their initial topic-relevance baseline MAP with about 11%.

The **University of Waterloo (UoW)** used a manually constructed list of 1336 subjective adjectives in document ranking. The top 1000 documents retrieved using BM25 were re-ranked based on the proximity of each query term instance to the subjective adjectives. Experiments were also conducted with different types of queries constructed from the topic titles: single terms and user-defined phrases, i.e. phrases enclosed in quotation marks by the user. Some improvements over the topic-relevance baseline were achieved (about 5.8% improvement) when the initial document set was retrieved using phrases, while the subjective adjective-based re-ranking was done using single terms. UoW concluded that subjective adjectives located close to any word from the query are useful indicators of the presence of opinions expressed about the query topic.

It is of interest to make some comments about the submitted official runs by some participating groups. The University of Illinois at Chicago (UIC) achieved the top scoring opinion finding run. How-

Group	Run	Fields	MAP	R-prec	b-Bref	P@10
UIC (Zhang)	uic1c	T	**0.4819**	0.5181	0.5484	**0.868**
UAmsterdam (deRijke)	uams07topic	T	0.4741	**0.523**	**0.5702**	0.762
FudanU (Wu)	FDUTisdOpSVM	T	0.4714	0.4889	0.5432	0.654
IndianaU (Yang)	oqlr2fopt	TDN	0.4347	0.4653	0.5022	0.822
CAS (Liu)	Relevant	T	0.4302	0.4949	0.5658	0.662
DalianU (Yang)	DUTRun2	T	0.4247	0.4750	0.5164	0.784
UGlasgow (Ounis)	uogBOPFProxW	T	0.4160	0.4436	0.4618	0.720
UNeuchatel (Savoy)	UniNEblog3	TD	0.4034	0.4296	0.4553	0.730
FIU (Netlab team)	FIUDDPH	TD	0.3907	0.4230	0.4692	0.714
UArkansas Littlerock (Bayrak)	UALR07BlogIU	T	0.3612	0.3975	0.4122	0.734
UWaterloo (Olga)	UWopinion3	T	0.3490	0.4040	0.4020	0.68
Zhejiangu (Qiu)	EAGLE2	T	0.3409	0.3809	0.3992	0.644
CAS (NLPR-IACAS)	NLPRTD	TD	0.3373	0.3804	0.3894	0.586
KobeU (Eguchi)	KobePrMIR01	T	0.3292	0.3655	0.3852	0.606
BUPT (Weiran)	prisOpnBasic	T	0.3267	0.3633	0.3735	0.684
NTU (Chen)	NTUManual	T	0.3051	0.3309	0.3631	0.582
RGU (Mukras)	rgu0	T	0.2798	0.3533	0.3651	0.560
KobeU (Seki)	Ku	T	0.2590	0.3357	0.3503	0.476
UBuffalo (Ruiz)	UB1	TDN	0.2421	0.2818	0.2956	0.484
Wuhan (Lu)	NOOPWHU1	T	0.0016	0.0111	0.0100	0.02

Table 7: Topic-relevance results: run from each of the 20 groups with the best topic-relevance MAP, sorted by MAP. The best in each column is highlighted.

ever, they did not submit the compulsory topic-relevance baseline. Therefore, it is difficult to assess the usefulness of their opinion finding features. Nevertheless, UIC's retrieval system contained two sub-systems. The opinion retrieval system (ORS), which was modified from the TREC 2006 version, and was used for the main task and a polarity classification system (PCS), which was newly designed for the polarity subtask. UIC experimented with a single query-independent SVM classifier and tested a spam detection module.

The runs submitted by the University of Amsterdam (UvA) raise a few interesting issues. While they had a strongly performing topic-relevance baseline run (see run uams07topic in Table 3), their used opinion finding features do not appear to be useful. UvA used the opinion finding task to compare the performance of an Indri implementation to their own mixture model. The mixture model combines different components of blog posts (e.g., headings, title, body) and assigns weights to these components based on tests on the TREC 2006 topics. Of both the baselines, the Indri system performed markedly better. To achieve better topical results, external (query) expansion on the AQUAINT-2 news corpus was performed. This expansion improves the performance of the Indri implementation, but hurts the mixture model. For opinion finding, UvA experimented with document priors in the mixture model based on either opinionated lexicons or the number of comments. The latter opinion finding features have not improved their opinion finding performance, markedly hurting their strongly performing uams07topic topic-relevance baseline run. In particular, run uams07topic is the 2nd top scoring title-only opinion finding run of the track, despite not using any opinion detection approach, suggesting that a strong retrieval baseline can do very well on the opinion finding task.

Interestingly, the Netlab team (FIU) used an approach that is very similar to the word list-based detection approach deployed by UoG, although developed separately. FIU used the DFR models, i.e. PL2 and the parameter free DPH, to assign both topic and opinion scores. A fully automatic and weighted dictionary was generated from TREC 2006's opinion finding relevance data. This dictionary was filtered and then submitted as a query to the Terrier search engine to get an initial query-independent opinion score of all retrieved documents. Ranking is done in two passages: a first topical-opinion ranking is obtained from the query-independent opinion score divided by the content rank, then the final topical-opinion ranking is established from the content score divided by the previous topical-opinion rank. Since FIU updated the final ranks but not the final topical-opinion scores in the re-ranking, trec_eval reported the same performance for all their official submitted runs. However, using the Terrier evaluation tool, which instead evaluates runs by ranks and not by scores, they show that FIU's opinion finding approach is actually effective. Indeed, their opinion finding run FIUIPL2 has about 17% improvement over their topic relevance baseline, an improvement in the same line as observed with UoG's wordlist-based approach, and expected given the similarities of the two groups's approaches.

3.6 Summary of Opinion Finding Task

The additional requirement that each participating group submits a compulsory topic-relevance baseline run allowed us to draw more conclusions on those opinion detection approaches that have worked and those that have not, providing additional insights for future work.

The overall opinion finding performance of the participating groups this year was markedly higher than the one observed for the TREC 2006 topics set. However, it is difficult to assess whether this increase in performance is due to the better deployed opinion finding systems and techniques or whether it is due to the difficulty level of the topics set. Answering this question requires running this year's systems on last year's topics.

Finally, similar to last year's conclusion, there appears to be no strong evidence that spam was a major hindrance to the retrieval performance of the participating groups.

4. BLOG DISTILLATION (FEED SEARCH) TASK

The blog distillation (feed search) task is a new task in the TREC 2007 Blog track, which was the result of the discussion that followed the introduction of the open task in TREC 2006. The task

Group	Run	Spam@10	Spam@all	BadMAP *10^-5
UIC (Zhang)	uic1c	**0.56**	**33.86**	**0.0**
UAmsterdam (deRijke)	uams07topic	0.92	104.14	2.8
UGlasgow (Ounis)	uogBOPFProxW	1.24	126.32	10.8
DalianU (Yang)	DUTRun2	0.74	55.66	3.0
FudanU (Wu)	FDUTOSVMSem	0.98	59.50	2.2
CAS (Liu)	Relevant	1.34	75.66	2.2
UArkansas Little Rock (Bayrak)	UALR07BlogIU	0.88	121.74	10.2
IndianaU (Yang)	oqsnr2opt	0.98	181.20	13.0
UNeuchatel (Savoy)	UniNEblog1	1.18	139.18	12.2
FIU (Netlab team)	FIUbPL2	1.42	131.98	11.8
UWaterloo (Olga)	UWopinion3	1.16	75.88	7.2
Zhejiangu (Qiu)	EAGLE1	1.24	121.74	9.6
CAS (NLPR-IACAS)	NLPRPST	1.22	124.68	8.4
BUPT (Weiran)	prisOpnBasic	1.32	80.22	7.2
KobeU (Eguchi)	KobePrMIR01	1.54	157.82	13.4
NTU (Chen)	NTUAutoOp	0.94	161.70	15.0
KobeU (Seki)	Ku	2.12	153.42	10.6
RGU (Mukras)	rgu0	1.30	86.30	5.6
UBuffalo (Ruiz)	UB2	4.92	86.44	4.2
Wuhan (Lu)	NOOPWHU1	1.56	101.96	4.8

Table 10: Spam measures for runs from Table 3, in the order given. Spam@10 is the mean number of spam posts in the top 10 ranked documents for each topic, Spam@all is the mean number of spam posts retrieved 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. For all measures, lower means the system was better at not retrieving spam documents. The best in each column is highlighted.

Group	Run	R-Acc	A@10	A@1000
UIC (Zhang)	uic75cpnm	**0.2295**	**0.3700**	**0.0493**
UAmsterdam (de Rijke)	uams07ipolt	0.1827	0.2640	0.0418
IndianaU (Yang)	oqsnr2optP	0.1799	0.2800	0.0401
DalianU (Yang)	DUTRun2P	0.1721	0.3080	0.0406
Zhejiangu (Qiu)	EAGLE2P	0.1510	0.2380	0.0427
UGlasgow (Ounis)	uogBOPFPol	0.1460	0.2020	0.0397
NTU (Chen)	NTUAutoOpP	0.0967	0.1860	0.0296
CAS (Liu)	DrapStmSub	0.0818	0.1060	0.0243
BUPT (Weiran)	pUB21	0.0418	0.0340	0.0148
Wuhan (Lu)	OTPSWHU102	0.0032	0.0040	0.0010

Table 12: Best polarity run for each group, in terms of R-accuracy. Each polarity runs corresponds to an automatic title-only opinion finding run. The best in each column is highlighted. Not all groups submitted polarity runs corresponding to automatic title-only opinion finding runs.

focuses on an interesting feature of the blogs, namely the fact that feeds are aggregates of blog posts.

4.1 Motivations

Blog search users often wish to identify blogs (i.e. feeds) about a given topic, which they can subscribe to and read on a regular basis. This user task is most often manifested in two scenarios:

- Filtering: The user subscribes to a repeating search in their RSS reader.
- Distillation: The user searches for blogs with a recurring central interest, and then adds these to their RSS reader.

For TREC 2007, the latter scenario was investigated i.e. blog distillation, which is a feed search task. The blog distillation task can be summarised as *Find me a blog with a principle, recurring interest in X*. For a given target *X*, systems should suggest feeds that are principally devoted to X over the timespan of the feed, and would be recommended to subscribe to as an interesting feed about *X* (i.e. a user may be interested in adding it to their RSS reader). This task is particularly interesting for the following reasons:

- A similar (yet-different) task has been investigated in the Enterprise track (Expert Search) in a smaller setting (around 1000 candidate experts on the W3C collection). For blog distillation, the Blog06 corpus contains around 100k blogs, and is a Web-like setting (with anchor text, linkage, spam, etc).
- A Topic distillation task was run in the Web track. In Topic distillation, site relevance was defined as (i) it is principally devoted to the topic, (ii) it provides credible information on the topic, and (iii) it is not part of a larger site also principally devoted to the topic.

While the definition of blog distillation as explained above is different, the idea is to provide the users with the key blogs about

Group	Run	Fields	R-Acc	A@10	A@1000
UIC (Zhang)	uic75cpnm	T	**0.2295**	**0.3700**	**0.0493**
IndianaU (Yang)	oqlr2f2optP	TDN	0.1941	0.3080	0.0438
UAmsterdam (de Rijke)	uams07ipolt	T	0.1827	0.2640	0.0418
DalianU (Yang)	DUTRun2P	T	0.1721	0.3080	0.0406
Zhejiangu (Qiu)	EAGLE2P	T	0.1510	0.2380	0.0427
UGlasgow (Ounis)	uogBOPFPol	T	0.1460	0.2020	0.0397
NTU (Chen)	NTUManualOpP	T	0.1161	0.2300	0.0348
CAS (Liu)	DrapStmSub	T	0.0818	0.1060	0.0243
BUPT (Weiran)	prisPolC2	T	0.0726	0.2020	0.0124
UBuffalo (Ruiz)	pUB11	TDN	0.0671	0.1000	0.0195
Wuhan (Lu)	OTPSWHU102	T	0.0032	0.0040	0.0010

Table 13: Best polarity run for each group, in terms of R-accuracy, regardless of the topic length. The best in each column is highlighted. Not all groups submitted polarity runs.

```
<top>
  <num> Number: 994 </num>

  <title> formula f1 </title>

  <desc> Description:
Blogs with interest in the formula
one (f1) motor racing, perhaps with
driver news, team news, or event
news.
</desc>

  <narr> Narrative:
  Relevant blogs will contain news
and analysis from the Formula f1
motor racing circuit. Blogs with
documents not in English are not
relevant.
</narr>
  </top>
```

Figure 5: Blog track 2007, blog distillation task, topic 994.

a given target. Note that point (iii) from the definition of the Web track Topic distillation task is not applicable in a blog setting.

4.2 Topics and Relevance Judgments

For the purposes of the blog distillation task, the retrieval document units are documents from the feeds component of the Blog06 collection. However, similar to the opinion finding task, the participating groups were free to use any other component of the Blog06 test collection in their submitted runs.

The topics for the blog distillation were created and judged by the participating groups. Each participating group has been asked to provide 6 or 7 topics along with some relevant feeds. A standard search system for documents on the Blog06 collection using the Terrier search engine [3] was provided by the University of Glasgow to help the participating groups in creating their blog distillation topics. The system displays the corresponding feed for each returned document (i.e. blog post), as well as all the documents for a given feed. Eight groups contributed each 5 to 7 topics. 45 topics were finally chosen by NIST from the proposed set of topics. A sample blog distillation topic is shown in Figure 5.

Overall, 9 groups submitted runs and agreed to help in their relevance judgments. Once runs were submitted, NIST formed pool and sent them to the University of Glasgow, where the community assessment system was hosted. The community judgments system interface was ported directly from the TREC Enterprise judgment system for expert search task developed by Soboroff et al. [4].

Participants were allowed to submit up to 4 runs, including a compulsory title-only run. Similar to the opinion finding task, the participants were asked to prioritise runs, in order to define which of their runs would be pooled. Each run has feeds ranked by their likelihood of having a principle (recurring) interest in the topic. Given the number of feeds in the collection (just over 100k feeds), each submitted run consisted of up to 100 feeds for each topic. A pool has then been formed by NIST from the 32 submitted runs, using the two highest-priority runs per group, pooled to depth 50.

For the assessment of the relevance of a feed, the assessors were asked to browse some of the documents of the feed, and then make a judgment on whether the feed has a recurring principle interest in the topic area. These guidelines are intentionally vague. A question that may arise is the number of documents (i.e. posts) that have to be read by the assessor for a given feed. Since there is no straightforward answer to this question, we decided to suggest that the assessors read enough documents of the feed such that they are certain that the feed has a more than passing interest in the topic area, and that they would be interested in subscribing to the feed in their RSS reader if they were interested in the topic area.

4.3 Overview of Results

The blog distillation task is another articulation of real user tasks in adhoc search behaviour on the blogosphere. Therefore, we use mean average precision (MAP) as the main metric for the evaluation of the retrieval performance of the submitted runs. In addition, we also report R-Precision (R-Prec), binary Preference (bPref), and Precision at 10 documents (P@10).

All submitted runs were automatic. Table 14 provides the average best, median and worst MAP measures for each topic, across all submitted 32 runs. Figure 6 shows the distribution of the number of relevant feeds per topic in the pooled feeds, sorted in decreasing order. In particular, there appears to be a wide variance in the number of relevant feeds across the used 45 topics, with topics having as many as 153 relevant feeds (e.g. “christmas” (968) or “music” (978)), while other having as few as 5 relevant feeds (e.g. “Violence in Sudan” (964) or “machine learning” (982)).

Table 15 shows the best-scoring automatic title-only run from each participating group in terms of MAP, and sorted in decreasing order. Table 16 shows the best run from each group, regardless of the topic length used. Note that most of the 32 submitted runs were title-only runs. Indeed, there were 25 submitted runs using the title field only, 3 submitted runs used the title, description and narrative

	MAP
Best	0.4671
Median	0.2035
Worst	0.0006

Table 14: Best, median and worst MAP measures for the 32 submitted runs to the blog distillation task.

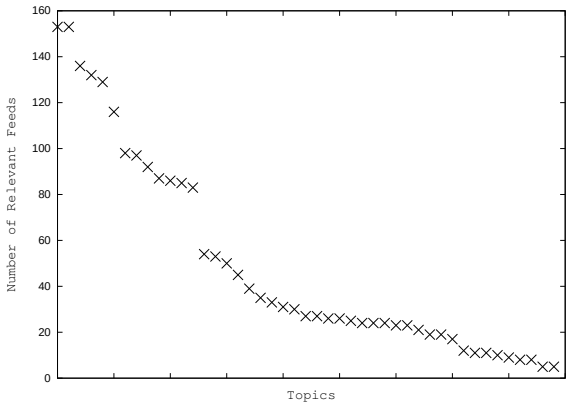


Figure 6: Distribution of number of relevant feeds per topic

Relevance Scale	Nbr. of Splogs
Not Relevant	2935
Relevant	255
(Total)	3190

Table 17: Occurrences of presumed splogs in the blog distillation task pool.

fields, 2 submitted runs used the title and description fields, and 2 submitted runs used the description field only. All the 10 best submitted blog distillation runs but one are title-only runs. Given the rather small number of submitted runs using long queries, it is difficult to draw conclusions as to whether the description and narrative fields of the topics might be helpful in the blog distillation task.

We examined whether the participating systems in the blog distillation task had been affected by spam, i.e. how many splog feeds have infiltrated the pool. Table 17 shows the breakdown of the feed distillation pool in terms of splog feeds. Moreover, Table 18 shows the extent to which the 17,958 presumed splogs have infiltrated the submitted runs. We use the mean number of splog documents in the top 10 ranked documents (denoted Spam@10), in the retrieved documents (Spam@all), and finally BadMAP, which is the Mean Average Precision when the splog feeds are treated as the relevant set. Run UMaTiPCSwGR from UMass appears to be overall the least susceptible to splog feeds. On the contrary, run TDWHU200 was one of the most affected runs by splog feeds.

Similar to the analysis performed in Section 3.3, to see if runs that retrieved less splog feeds were more likely to be high performing systems or low performing systems, we correlated the ranking of submitted runs by BadMAP, correlating this with blog distillation MAP. For this task, a weak correlation was exhibited ($\rho = -0.193, \tau = -0.157$), showing some evidence that systems which did remove splogs were likely to have a higher retrieval performance.

4.4 Participant Approaches

There were a wide range of deployed indexing and retrieval approaches for the blog distillation task. The exploratory nature of most of the used techniques characterises the novelty of the task and its interesting underlying features. The main features of the submitted runs are summarised below:

Indexing Two types of indexes have been used. Three groups created an index using the Feeds component of the Blog06 collection, namely Carnegie Mellon University (CMU), the University of Texas, and the University of Wuhan. The rest of the groups only indexed the Permalinks component of the collection. Interestingly, CMU, the top performing group, experimented with both types of index, and concluded that an index based on the Feeds component of the Blog06 collection leads to a better retrieval performance on this task.

Retrieval Many groups approached the blog distillation task by connecting the task to other existing search tasks. For example, the University of Glasgow (UoG) explored the connection of blog distillation to the expert finding task of the Enterprise track, adapting their Voting Model paradigm to feed search. The University of Massachusetts looked at the blog distillation task as a resource selection problem in distributed search. Most of the groups that used an index based on Permalinks, have proposed various techniques to aggregate the scores of blog posts into a score for their composing feed. For the purposes of document retrieval, a range of document weighting models such as Language Modelling approaches and Divergence From Randomness models were used. Some groups have also experimented with classical information retrieval techniques, namely query expansion (e.g. CMU) and proximity search (e.g. UoG).

In the following, we provide a detailed description of the methods used by the top 3 performing groups in the blog distillation task:

Carnegie Mellon University (CMU) explored two indexing strategies, namely a large-document model (feed retrieval) and a small-document model (entry or blog post retrieval). Under the large-document model, feeds were treated as the unit of retrieval. Under the small-document model, the blog posts were treated as the unit of retrieval and aggregated to produce a final ranking of feeds. They found that the large-document approach outperformed the small-document approach on average. CMU also experimented with a query expansion method using the link structure and link text found within an external resource, namely the Wikipedia. CMU found that the used Wikipedia-based query expansion approach improves results under both the large- and small-document models.

The **University of Glasgow (UoG)** only indexed the Permalink component of the Blog06 collection. They investigated the connections between the blog distillation task and the expert search task. UoG adapted their Voting Model paradigm for Expert Search, by ranking feeds according to the number of on-topic posts each feed has (number of votes), and the extent to which the posts are about the topic area (strength of votes) - these two sources of evidence about the interests of each blogger were combined using the expCombMNZ voting technique. Posts are ranked using the PL2F Divergence From Randomness (DFR) field-based weighting model. They found that the additional use of a DFR-based term proximity model improves the topicality of the underlying ranking of blog posts, leading to a more accurate aggregated ranking of blog posts and a better feed search performance.

The **University of Massachusetts (UMass)** used language modelling approaches. UMass used the Permalink component of the Blog06 test collection for indexing. UMass looked at this task as a resource selection problem in distributed information retrieval,

Group	Run	MAP	R-prec	b-Bref	P@10	MRR
CMU (Callan)	CMUfeedW	**0.3695**	**0.4245**	**0.3861**	**0.5356**	0.7537
UGlasgow (Ounis)	uogBDFeMNZP	0.2923	0.3654	0.3210	0.5311	0.7834
UMass (Allen)	UMaTiPCSwGR	0.2529	0.3334	0.2902	0.5111	**0.8093**
KobeU (Seki)	kudsn	0.2420	0.3148	0.2714	0.4622	0.7605
DalianU (Yang)	DUTDRun1	0.2285	0.3105	0.2768	0.3711	0.5813
UTexas-Austin (Efron)	utblnrr	0.2197	0.3100	0.2649	0.4511	0.7245
UAmsterdam (deRijke)	uams07bdtblm	0.1605	0.2346	0.1820	0.3067	0.6320
WuhanU (Lu)	TDWHU200	0.0135	0.0419	0.0297	0.0578	0.1386

Table 15: Blog distillation results: the automatic title-only run from each of 8 groups with the best MAP, sorted by MAP. Note that 1 group (UBerlin) did not submit a title-only run. The best in each column is highlighted.

Group	Run	Fields	MAP	R-prec	b-Bref	P@10	MRR
CMU (Callan)	CMUfeedW	T	**0.3695**	**0.4245**	**0.3861**	**0.5356**	0.7537
UGlasgow (Ounis)	uogBDFeMNZP	T	0.2923	0.3654	0.3210	0.5311	0.7834
UMass (Allen)	UMaTDPCSwGR	TD	0.2741	0.3356	0.3027	**0.5356**	**0.8407**
KobeU (Seki)	kudsn	T	0.2420	0.3148	0.2714	0.4622	0.7605
DalianU (Yang)	DUTDRun4	TDN	0.2399	0.3126	0.2740	0.4378	0.7337
UTexas-Austin (Efron)	utblnrr	T	0.2197	0.3100	0.2649	0.4511	0.7245
UAmsterdam (deRijke)	uams07bdtblm	T	0.1605	0.2346	0.1820	0.3067	0.6320
UBerlin (Neubauer)	ADABoostM1	TDN	0.0176	0.0468	0.0330	0.0978	0.2881
WuhanU (Lu)	TDWHU200	T	0.0135	0.0419	0.0297	0.0578	0.1386

Table 16: Blog distillation results: one run from each of 9 groups with the best MAP, sorted by MAP. The best in each column is highlighted.

since each feed can be considered as a collection composed of blog posts. The most critical issue of resource selection is how a collection is represented. UMass applied two approaches for representation in this task. Further, since blogs which address many general and shallow topics are unlikely to be relevant in this task, UMass introduced an approach to penalise such blogs, and found that this improves the retrieval effectiveness.

Other approaches used by the participating groups included the investigation of blog specific approaches such as time-based priors and splog detection and filtering, or retrieval models variants to search from a feeds-based index. The University of Amsterdam (UvA) experimented with time-based priors. Their suggested idea is that more recent posts reflect better the current interest of a blogger. Results show that time-based priors, which order the feeds based on the score of the most relevant post from a feed, improve slightly over the baseline run. UvA also experimented with a relevant posts count, where for every feed the ratio of relevant posts to all posts in a feed is calculated and this score is combined with the feed relevance score from the baseline run. Results show that this has markedly decreased performance, suggesting that the combination parameters were not appropriate.

Kobe University (Seki et al.) experimented with splog detection, and filtering of non-English documents. Interestingly, their baseline is built by computing the similarity scores between a query and the posts included in the feed. They plotted a line for each blog site with the x-axis being the (normalised) post date and the y-axis being the computed similarity. The feeds are then ranked according to the descending order of the surface area under the plotted line. The intuition behind the proposed algorithm is that a relevant feed would frequently mention a given topic, and will constantly have a high similarity with the topic (query), resulting in a large surface area under the line of similarity scores. They found that filtering splogs and non-English documents improves their baseline.

Finally, the University of Texas' School of Information (UT) used a retrieval strategy based on a variant of the Kullback-Leibler (KL) divergence model. Given a query q the UT system derives a score for each feed f in the corpus by the negative KL-divergence between the query language model and the language model for f. The effectiveness of the proposed approach cannot be assessed without an experimental baseline.

4.5 Summary of Blog Distillation Task

The blog distillation task was a new task in TREC 2007. Overall, some of the deployed retrieval approaches achieved reasonable retrieval performances. One of the issues that might need to be further investigated in this task is whether it is beneficial to use the Feeds component of the Blog06 collection, instead of or in addition to the Permalinks component.

There was a wide variance in the distribution of relevant feeds in the used 45 topics, suggesting that the guidelines for the topic creation and assessments still require tightening for future iterations of this task. However, the task, as exemplified by the exploratory nature of the participants runs, promises much research in the future.

5. CONCLUSIONS

The TREC 2007 Blog track included two main tasks, namely the opinion finding and Blog distillation (aka feed search) tasks, which we believe are good articulations of real user tasks in adhoc search behaviour on the blogosphere. The used tasks address two interesting components of blogs: the feed itself and its constituent blog posts and their corresponding comments. As a consequence, a new topics set has been created for the opinion finding task, and a new test collection has been created for the Blog distillation task, therefore contributing to the creation of reusable resources for supporting research into blog search.

Much remains to be learned about opinion finding, even though the runs submitted this year show that some participants have been successful in proposing new opinion detection techniques, which show some marked improvements on the respective topic-relevance baseline. Indeed, this year's findings also consolidate the findings

Group	Run	Spam@10	Spam@all	BadMAP *$10^{-}5$
CMU (Callan)	CMUfeedW	2.8	22.5	48.2
UGlasgow (Ounis)	uogBDFeMNZP	2.2	22.4	28.0
UMass (Allen)	UMaTiPCSwGR	**0.6**	**3.1**	**3.1**
KobeU (Seki)	kudsn	1.5	9.2	10.0
DalianU (Yang)	DUTDRun1	3.6	21.6	56.2
UTexas-Austin (Efron)	utblnrr	2.0	15.5	23.7
UTexas-Austin (Efron)	utlc	2.1	13.44	19.6
UAmsterdam (deRijke)	uams07bdtblm	1.9	13.7	26.0
WuhanU (Lu)	TDWHU200	3.1	159.1	184.0

Table 18: Spam measures for runs from Table 15, in the order given. Spam@10 is the mean number of splog feeds in the top 10 ranked documents for each topic, Spam@all is the mean number of splog feeds retrieved for each topic. BadMAP is the Mean Average Precision when the splog feeds are treated as the relevant set. This shows when spam feeds are retrieved at high ranks. For all measures, lower means the system was better at not retrieving splogs.

of the previous Blog track 2006. In particular, a good performance in opinion finding is strongly dominated by its underlying topic-relevance baseline (i.e. opinion-finding MAP and topic-relevance MAP are very highly correlated). Indeed, a strongly performing topic-relevance baseline can still perform extremely well in opinion finding, as exemplified by the University of Amsterdam's submitted topic-relevance baseline. One possible methodology to have a better understanding of the deployed opinion detection techniques is to use a common and strong topic-relevance baseline for all participating groups.

For the polarity subtask, the overall performances of the participating groups are rather average, suggesting that the task of detecting the polarity of an opinion is still an open problem, which requires further research. We believe that polarity detection should be a more integral part of the opinion finding task, and not evaluated as in classification task-like manner. For future iterations of the opinion finding task, we believe that a better integration of the polarity component would involve creating a balanced number of topics, which explicitly specify whether they require positive or negative opinions to be retrieved. Evaluation can then be carried out in a more straightforward adhoc manner.

The Blog distillation task seems to have generated some very promising and interesting retrieval techniques. We plan to run the task again for 2008, in a similar fashion, but with clearer guidelines for the creation of the topics. This will provide further insights on the most effective techniques for this task.

Acknowledgments

The description of systems runs are based on paragraphs contributed by the participating groups. Thanks are also due to participants in the blog distillation task for creating and assessing topics. In particular all groups who submitted runs to the blog distillation task have participated in the relevance judging. Finally, we are grateful to Arjen de Vries for providing his assessment system, and to David Hannah for adapting it to the blog distillation task.

6. REFERENCES

[1] C. Macdonald and I. 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] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, I. Soboroff. Overview of TREC-2006 Blog track. In *Proceedings of TREC-2006*, Gaithersburg, USA, 2007.

[3] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of OSIR'2006 Workshop*, Seattle, USA, 2006

[4] I. Soboroff, A. de Vries, and N. Craswell. Overview of TREC-2006 Enterprise track. In *Proceedings of TREC-2006*, Gaithersburg, USA, 2006.