Overview of the TREC-2008 Blog Track

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

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

1. INTRODUCTION

The Blog track explores the information seeking behaviour in the blogosphere. The track was introduced in 2006 [1], with a main pilot search task, namely the opinion-finding task. In TREC 2007 [2], the track investigated two main tasks inspired by the analysis of a commercial blog-search query log: the opinion-finding task (i.e. "What do people think about X?") and the blog distillation task (i.e. "Find me a blog with a principal, recurring interest in X."). In addition, the Blog 2007 track investigated a natural extension to the opinion-finding task, namely the polarity task (i.e. "Find me positive or negative opinionated posts about X."). All tasks thus far investigated in the Blog track have used the so-called Blogs06 collection, which was created by the University of Glasgow [3]. The Blogs06 collection was crawled over an 11-week period from 6th December 2005 until the 21st February 2006. The collection is 148GB in size, consisting of 38.6GB of feeds, 88.8GB of permalink documents, and 28.8GB of homepages.

For TREC 2008, the track continued using the Blogs06 collection. It also continued investigating the opinion-finding, polarity, and blog distillation tasks. In addition, the Blog track 2008 introduced a baseline blog post retrieval task (i.e. "Find me blog posts about X."), to encourage participants to study the impact of their opinion-finding techniques across different underlying topic-relevance baselines. As a consequence, following our conclusions from both the TREC 2006 and the Blog 2007 tracks, we structured the Blog track 2008 around four tasks:

(1) Baseline adhoc (blog post) retrieval task;

(2) Opinion-finding (blog post) retrieval task;

(3) Polarity opinion-finding (blog post) retrieval task; and

(4) Blog (feed) distillation task.

The track has seen an increased level of participation over the years from 17 groups in 2006, to 24 groups in 2007 (20 participants in the opinion-finding task, 11 in the polarity task, and 9 in the blog distillation task). In TREC 2008, 20 groups submitted runs to the baseline task, 19 groups submitted runs to the opinion-finding task, 16 groups submitted runs to the polarity task, and 12 groups submitted runs to the blog distillation task.

The remainder of this paper is structured as follows. Section 2 describes the baseline and opinion-finding tasks, providing an overview of the submitted runs, as well as a summary of the main effective techniques used by the participating groups. Section 3 describes the polarity task, and the main obtained results by the participating groups. Section 4 describes the blog search (blog distillation) task, and summarises the results of the runs and the main effective approaches deployed by the participating groups. We provide concluding remarks in Section 5.

2. BASELINE AND OPINION-FINDING TASKS

2.1 Tasks and Topics

The opinion-finding task addresses a search scenario where a user aims to uncover what the bloggers are saying about X. Roughly speaking, the user's intention is to "take the pulse of the blogosphere" on a topic X. The task has been running in TREC since the Blog track inception in 2006 [1]. One of the lessons learnt from Blog tracks of TREC 2006 & TREC 2007 is that a good performance in opinion-finding is strongly dominated by the underlying document ranking performance (topic-relevance baseline), where the system's aim is to retrieve as many relevant documents as possible regardless of their opinionated nature [1, 2]. In addition, while some participants were able to show a marked increase in performance when using opinion detection features on top of good topic-relevance baselines, other groups did not manage to improve their baselines. In a recent study, we showed that some stronger topic-relevance baselines could not be improved even by applying the most effective opinion-finding approaches proposed in TREC 2007 [6].

As a consequence, to allow the further study of the performance of a specific opinion-finding technique across a range of different topic-relevance baseline systems, we introduced a two-stage submission procedure for the opinion-finding task. In the first stage (baseline adhoc retrieval task), the participating groups were asked to submit their topic-relevance baselines. Five submitted topic-relevance baseline runs were then selected by TREC as the "standard baselines" and made available to the participating groups. These standard baselines use a variety of different retrieval approaches, and have varying retrieval effectiveness. More specifically, they were selected based on their high topic-relevance and opinion-finding performances on the TREC 2006 and TREC 2007 old topics. Table 1 summarises the five provided standard baseline runs.

In the second phase (opinion-finding retrieval task), the participating groups were encouraged to apply their opinion-finding techniques on their own baselines and on as many standard baselines as possible. The idea was to provide the participating groups with an experimental setting where they could assess the impact of their opinion-finding techniques across a range of different topic-relevance baselines or independently of their own baselines. Through this experiment, the Blog track 2008 also aimed to draw a better understanding of the most effective and stable opinion-finding techniques, by observing their performances on common standard topic-relevance baselines.

This experiment was made possible by the fact that most of the participating groups in both TREC 2006 and 2007 approached the opinion-finding task as a re-ranking problem [1, 2, 5]. In the first

Baseline	Run ID	Run type	Topics
baseline1	uicirwa	Automatic	Title-only
baseline2	DCUCDVPtdbl	Automatic	Title-desc
baseline3	UniNEBlog1	Automatic	Title-desc
baseline4	KLEPsgFeedTD	Automatic	Title-desc
baseline5	prisbm	Manual	Title-only

Table 1: Details of the five provided standard baselines.

stage, a group's system aims to find as many relevant documents as possible, regardless of their opinionated nature, while in the second stage, the system re-ranks those documents using some opinion detection techniques, and an appropriate combination of scores. For those participating groups that could not separate the topic-relevance and opinion-finding components, the submission guidelines were flexible enough to allow these groups to submit runs without the requirement of specifying a baseline run.

Since the commercial query logs used in TREC 2006 and 2007 have been running out of workable topics, for TREC 2008, the assessors were asked to create 50 new topics using the query logs as a source, but also by following their own ideas when browsing the collection. Groups were asked to submit their runs using the 50 new topics, as well as the 100 queries from the TREC 2006 and 2007 opinion-finding tasks. The idea was to draw conclusions about the difficulty of the query topics across the Blog track years as well as to provide the participating groups with an experimental setting allowing them to evaluate their training methods and re-ranking functions. In fact, our study in [6] shows that it is often necessary to train the used re-ranking function.

2.2 Assessments and Pools

Each submitted run consisted of the top 1000 retrieved documents for each topic. The retrieval units are the documents from the permalinks component of the Blogs06 test collection. 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. We used the same assessment procedure as defined in the TREC Blog tracks 2006 and 2007 [1, 2]. In particular, the assessment procedure had two levels. The first level assesses whether or not 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. The relevance assessments were conducted by NIST.

Groups were allowed to submit at most 2 baseline runs, including a compulsory automatic title-only run, and up to 4 opinion-finding runs using their own baselines, again including a compulsory automatic title-only run. In addition, groups could submit up to 4 runs using each of the 5 provided standard baselines. Hence, each group could submit up to 24 opinion-finding runs. TREC received 41 baseline runs from 20 groups, and 191 opinion-finding runs from 19 groups. Of the 191 submitted opinion-finding runs, all but two runs were automatic: run prisbm (baseline run) and run prisom1 (opinion-finding), which were both manual runs by the BUPT_pris_ group. Among the opinion-finding runs, 130 runs used one of the provided standard baselines, 12 runs had N/A for the baseline (i.e. their system does not separate topic-relevance from opinion-finding), and the other 49 runs used a baseline run from the corresponding group. For the 130 runs using one of the standard baselines, Table 2 shows the number of runs using each baseline type, including the breakdown per standard baseline. The baseline, opinion-finding, and polarity tasks shared the same pool. NIST pooled the top 100 documents of two opinion-finding and one polarity runs per group. If a group didn't have any opinion-finding or polarity runs, it only contributed runs from the task it did participate in.

Baseline	Number of submitted runs
baseline1	25
baseline2	24
baseline3	24
baseline4	30
baseline5	27
(own)	49
(N/A)	12
Total	191

Table 2: Breakdown of the baselines used by the submitted opinion-finding runs, including the five standard baselines. Own denotes when a run was based on a participating group's own baseline retrieval system, while N/A denotes when a participant's system did not submit separate topic-relevance and opinion-finding runs.

Relevance level	2006	2007	2008
Not Relevant	949.82	848.68	841.60
Relevant	167.22	103.74	58.76
Relevant, negative opinions	74.14	36.88	55.78
Relevant, mixed opinions	73.28	43.92	53.40
Relevant, positive opinions	83.18	59.20	66.76
Total	1347.64	1092.42	1076.30

Table 3: Average number of judged documents per topic in each of the considered relevance levels across years 2006-2008.

Table 3 shows a breakdown of the average pool size per topic, and the distribution of relevance assessment levels over the three years of the Blog track opinion-finding task. It is of note that the TREC 2006 pool had the largest size. On the other hand, the TREC 2007 topics were the least opinionated. Table 3 also shows that, on average, each of the three pools had roughly an equal number of negative and mixed opinionated documents, but slightly more positive opinionated documents, suggesting that, overall, bloggers had more positive opinions about the topics tackled by the three years of the track.

2.3 Results

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

Table 4 provides the average best, median, and worst MAP and P@10 measures for each topic, across all submitted 41 baseline runs. Table 5 provides the same measures across all submitted 191 opinion-finding runs. Note that the medians are calculated using the "lower medians" and using only the submitted runs for the given task[1]. In particular, it is of interest to note that the retrieval performances of the systems on the TREC 2006 topics were markedly lower than those obtained on the TREC 2007 and TREC 2008 topics both in terms of topic-relevance and opinion-finding,

[1]The TREC distributed opinion-finding medians for each topic are computed over all runs (baselines + opinion-finding runs, i.e. 191+41 = 232 runs). We consider that including the baseline runs in the computation of the medians would be inappropriate, as these baseline runs were not intended to retrieve opinionated documents. In addition, when the number of runs is even (e.g. 232), TREC computes the "upper median".

using both the MAP and P@10 evaluation measures. This suggests that the 2006 topics were slightly more difficult than those used in TREC 2007 and 2008. On the other hand, on average, the performances of the participating groups on the TREC 2007 topics dataset were markedly higher than those reported last year for the same dataset [2]. However, it is unclear whether this is due to the deployed systems having better retrieval approaches or to intensive training. Nevertheless, it is of note that the performances of the participating groups on the unseen TREC 2008 queries were higher than those observed in TREC 2007, while being overall comparable to the performances of the same (trained) systems on the TREC 2007 dataset. This might suggest that the TREC 2008 topics are the easiest. Table 6 provides the average best, median, and worst MAP and P@10 measures for each topic, across all 2006-2008 years (150 topics), for all submitted 41 baseline and 191 opinion-finding runs.

In the following, to limit the influence of the training that some participating groups might have performed on the TREC 2006 and TREC 2007 topics, we only present the results corresponding to the 50 TREC 2008 unseen queries. Table 7 shows the best-scoring baseline title-only automatic run for each group in terms of topic-relevance MAP, and sorted in decreasing order. R-Prec, bPref and P@10 measures are also reported. Table 8 shows the best baseline run from each group, in terms of topic-relevance MAP, regardless of the topic length used. All top ranked runs are title-only runs but one.

The top ranked group, KLE, deployed a passage-based retrieval language modelling approach. Other groups, such as UAms and UoGtr, used collection enrichment, by applying query expansion on external news corpora. In addition, UAms's run included the use of document priors based on credibility indicators such as spelling and capitalisation. UoGtr's run applied a Divergence From Randomness (DFR) term dependency model to boost documents where query terms appear in close proximity. The UIC group used a concept-based information retrieval system and phrasal search. The UniNE group merged two title-only runs based on a 2-Word indexing strategy: one run applies query expansion, while the second applies collection enrichment using Wikipedia. Tables 7 and 8 also report the opinion-finding MAP measures for these baseline runs. It is of note that the overall rankings of the 41 baseline systems on either the opinion-finding or topic-relevance measures are very similar, as stressed by the obtained high correlation coefficients, namely Spearman's $\rho = 0.9934$ and Kendall's $\tau = 0.9488$.

In TREC 2008, the participating groups were encouraged to apply their opinion-finding techniques on top of their own baselines, as well as on as many of the provided five standard baselines as possible. Table 9 shows the best-scoring opinion-finding run for each group in terms of opinion-finding MAP, regardless of the used baseline and the query type. Other metrics reported are R-Precision (R-Prec), binary Preference (bPref), and Precision at 10 (P@10). In the table, we also compare the opinion-finding MAP performance of the run to the opinion-finding MAP performance achieved by its underlying topic-relevance baseline. A relative MAP increase in performance indicates that the used opinion-finding features were useful. A relative MAP decrease in performance indicates that the deployed opinion-finding features did not help in retrieval (see column Δ MAP). It is interesting to note that the best two runs used a system that does not clearly separate the topic-relevance and the opinion-finding components. Table 10 shows the best-scoring opinion-finding run for each group in terms of opinion-finding MAP, when the group used one of its own submitted baseline runs, regardless of the query type.

Tables 9 and 10 show that several groups managed to improve the opinion-finding performance of their underlying topic-relevance baseline. However, the improvements are rather slim, especially when the used topic-relevance baseline is strong enough (e.g. run uams08-nlolsp using the strongly performing baseline run uamso8n1o1). On the other hand, run DUTIR08BRun4, which led to the highest improvement over the used baseline (31.60%), did not use the best baseline submitted by the corresponding group (see Tables 7 & 10).

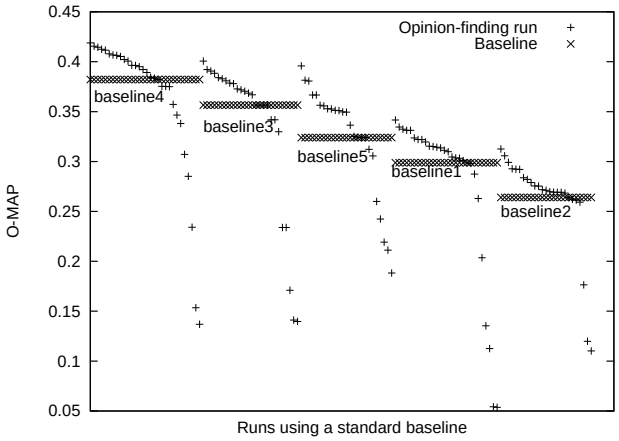

Figure 1: For each of the 130 opinion-finding task runs using a standard baseline, this figures shows the opinion-finding MAP (denoted O-MAP) of the opinion-finding task run compared to the opinion-finding MAP of the corresponding baseline. Ordering is by baseline run performance then opinion-finding run performance.

Among the five provided standard baselines, baseline4, (run KLE-PsgFeedTD), which used title and description topics, had the highest topic-relevance and opinion-finding MAP on the 50 new TREC 2008 queries. Table 11 shows the median of the opinion-finding runs using each of the standard baselines. According to Table 2, it is also the most frequently used one among the provided standard baselines. Table 12 shows the best performing opinion-finding run from each group, if and when the corresponding system used baseline4 as the baseline. In fact, putting apart those two runs that used a system that cannot separate the topic-relevance and opinion-finding components, the top 4 best runs in Table 9 all used baseline4 as their underlying baseline. This observation is further emphasised in Figure 1. For each opinion-finding task run using a standard baseline run, the figure shows how the opinion-finding MAP relates to the opinion-finding MAP of the corresponding baseline run. Indeed, most of the top runs used baseline4. However, there were also some approaches which did not perform well using this baseline.

Furthermore, we investigated the extent to which a given opinion-finding technique improved the opinion-finding MAP of all the 5 provided standard baselines. The more an opinion-finding technique consistently improves the opinion-finding retrieval performance of the 5 provided baselines, the more likely that it is effective. For a fair comparison of the opinion-finding techniques, we only considered the groups who attempted their opinion-finding techniques on all 5 provided standard baselines. Overall, 21 sets of runs using all five standard baselines were submitted by 8 groups. Table 13 shows the best opinion-finding approach from each of the 8 groups, ranked by the mean of their relative improvements over the five standard baselines (see column Mean Δ MAP). The mean of the opinion-finding performance of the corresponding run on the five standard baselines is also reported (see column Mean MAP). Table 13 shows that only three groups had opinion-finding approaches that seem to be effective across the five standard base-

	2006 (851-900)				2007 (901-950)				2008 (1001-1050)			
	MAP_{rel}	$P@10_{rel}$	MAP_{op}	$P@10_{op}$	MAP_{rel}	$P@10_{rel}$	MAP_{op}	$P@10_{op}$	MAP_{rel}	$P@10_{rel}$	MAP_{op}	$P@10_{op}$
median	0.3152	0.6800	0.2080	0.4220	0.3973	0.7280	0.2940	0.4880	0.3529	0.6960	0.2890	0.5700
best	0.5049	0.9440	0.3664	0.7580	0.6498	0.9600	0.4991	0.8000	0.5994	0.9140	0.5002	0.8260
worst	0.0242	0.0480	0.0131	0.0260	0.0532	0.0780	0.0281	0.0220	0.0381	0.0780	0.0284	0.0520

Table 4: Baseline runs: Best, median, and worst topic-relevance and opinion-finding MAP and P@10 measures of the 2008 participating groups across the three topic sets.

	2006 (851-900)				2007 (901-950)				2008 (1001-1050)			
	MAP_{rel}	$P@10_{rel}$	MAP_{op}	$P@10_{op}$	MAP_{rel}	$P@10_{rel}$	MAP_{op}	$P@10_{op}$	MAP_{rel}	$P@10_{rel}$	MAP_{op}	$P@10_{op}$
median	0.3408	0.7620	0.2549	0.5360	0.4407	0.8220	0.3552	0.6060	0.3819	0.7140	0.3291	0.6100
best	0.5747	0.9860	0.6456	0.9580	0.6965	0.9840	0.7626	0.9480	0.6279	0.9400	0.5610	0.8980
worst	0.0598	0.0340	0.0459	0.0140	0.0482	0.0100	0.0322	0.0040	0.0405	0.0060	0.0330	0.0020

Table 5: Opinion-finding runs: Best, median, and worst topic-relevance and opinion-finding MAP and P@10 measures of the 2008 participating groups across the three topic sets.

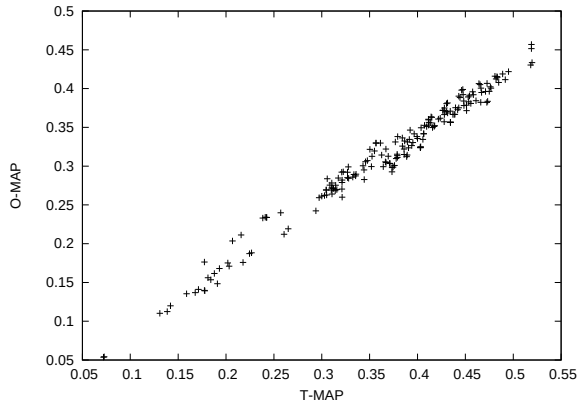

Figure 2: Scatter plot of opinion-finding MAP (O-MAP) against topic-relevance MAP (T-MAP) for all of the 191 submitted opinion-finding task runs.

lines: UIC_IR Group, KLE and UoGtr. Interestingly, from Table 13, we observe that Mean Δ MAP and Mean MAP are correlated, indicating that those opinion-finding techniques which on average do best are also the most stable across all five standard baselines.

Finally, for the 191 submitted opinion-finding runs, we computed the correlation between the opinion-finding MAP, and the topic-relevance MAP. The overall rankings of systems on both opinion-finding and topic-relevance measures are very similar, as stressed by the obtained high correlations, namely, Spearman's ρ=0.9862 and Kendall's τ=0.9054. Figure 2 shows a scatter plot of opinion-finding MAP against topic-relevance MAP, which confirms that the correlation is very high. Overall, similar to previous years, a good performance on the opinion-finding task is strongly dominated by a good performance on the underlying document retrieval task.

2.4 Participants Approaches

All the participating groups only indexed the permalinks component of the Blogs06 collection, with the exception of the THUIR group, which experimented with two indices: one based on the permalinks component of the Blogs06 collection and, for two submitted runs only, with an index based on both the permalinks and feeds components of the collection.

In terms of opinion-finding approaches, similar to the general trend in TREC 2006 and 2007 [1, 2], most of the submitted runs used a two-stage approach, where an initial set of relevant but not necessarily opinionated documents are re-ranked by taking into account various document opinion features. Only 12 runs out of the submitted 191 runs did not adopt this strategy, instead deploying a system that does not separate the topic-relevance component from the opinion-finding features.

We focus on those three opinion-finding approaches that were consistently effective across the five provided baselines as shown in Table 13. The approach uicop1bl1r, deployed by UIC_IR_Group, achieved the best average opinion-finding improvements over the five standard topic-relevance baselines (an average of 11.76% improvement). The UIC_IR_Group's opinion identification component uses an SVM classifier to distinguish subjective texts from objective texts, and determines whether each opinion in the subjective text is related to the query. Its effectiveness is enhanced by a concept abbreviation component, which attempts to recognize abbreviated query concepts in the vicinity of an opinion. The approach B1PsgOpinAZN, from the KLE group, used a lexicon-based approach. The opinion score of a given term is estimated using SentiWordNet and the Amazon review data. The opinionated level of a blog post is defined as the sum of opinion scores of terms within the post. The scores are normalised to take into account the length of the blog post. The KLE group used the Okapi's length normalisation component of BM25. Finally, the uogOP1PrintL opinion-finding approach, deployed by the UoGtr group, confirmed its effectiveness in the TREC Blog track 2007, by improving the opinion-finding performance of the five provided standard baselines by an average of 5.21%. Moreover, the UoGtr group enhanced their TREC 2007 dictionary-based approach [8] by automatically building an internal opinion dictionary from the collection itself. This approach measures the opinionated discriminability of each term in the dictionary using an information theoretic divergence measure based on the relevance assessments of previous opinion-finding tasks. In addition, UoGtr experimented with a novel method to measure the informativeness of the query terms occurring in a close proximity to opinionated sentences.

3. POLARITY TASK

One of the conclusions from the TREC 2007 Blog track is that the polarity detection task should be a more integral part of the opinion-finding task [2]. In particular, instead of being defined as a classification task where the system merely identifies the opinion direction (positive, negative, or mixed) of a blog post, the task has been redefined to simulate a user search scenario where the system would retrieve both the positive and negative opinionated

	Baseline runs				Opinion-finding runs			
	MAP_{rel}	$P@10_{rel}$	MAP_{op}	$P@10_{op}$	MAP_{rel}	$P@10_{rel}$	MAP_{op}	$P@10_{op}$
median	0.3551	0.7013	0.2636	0.4933	0.3878	0.7660	0.3131	0.5840
best	0.5847	0.9393	0.4552	0.7947	0.6330	0.9700	0.6564	0.9347
worst	0.0385	0.0680	0.0232	0.0333	0.0495	0.0167	0.0370	0.0067

Table 6: Baseline and opinion-finding runs over all 150 topics.

		Topic-Relevance					Opinion-Finding				
Group	Run	MAP	R-prec	bPref	P@10	MRR	MAP	R-prec	bPref	P@10	MRR
KLE	KLEPsgFeedT	0.4954	0.5150	0.5364	0.7920	0.9058	0.4052	0.4366	0.4314	0.6440	0.8184
UAms_De_Rijke	uams08n1o1	0.4644	0.4867	0.5034	0.7620	0.8892	0.3797	0.4176	0.4117	0.6620	0.8052
UIC_IR_Group	uicirnoa	0.4403	0.4804	0.5062	0.7700	0.8667	0.3438	0.3956	0.3929	0.5880	0.7480
UniNE	UniNEBlog2	0.4283	0.4551	0.4659	0.6580	0.8482	0.3537	0.3781	0.3676	0.5620	0.7963
UoGtr	uogBLProxCE	0.4219	0.4548	0.4481	0.7060	0.8228	0.3531	0.3840	0.3646	0.6100	0.7723
THUIR	THUrelTwpmf	0.4067	0.4565	0.4625	0.6940	0.8263	0.3313	0.3942	0.3749	0.5900	0.7487
BUPT_pris_	prisba	0.4065	0.4506	0.4561	0.6780	0.8290	0.3346	0.3876	0.3684	0.5580	0.7456
DUTIR	DUT08BRun1	0.3617	0.4188	0.4345	0.6540	0.7633	0.2974	0.3586	0.3598	0.5420	0.7204
iitkgp	IITKGPNOSPAM	0.3598	0.4090	0.4394	0.7400	0.8817	0.2988	0.3664	0.3642	0.5720	0.7955
IU-SLIS	wdoqsBase	0.3431	0.3918	0.4001	0.7280	0.8636	0.2818	0.3367	0.3215	0.5900	0.7551
UWaterlooEng	UWBase2	0.3309	0.3824	0.3875	0.6380	0.8127	0.2753	0.3391	0.3249	0.5160	0.7254
aic-dcu	DCUCDVPtbl	0.3303	0.3671	0.3601	0.6520	0.7783	0.2875	0.3280	0.3089	0.5560	0.7066
UIUC	UIUCb08uwTtl	0.3240	0.3766	0.3771	0.6800	0.8223	0.2723	0.3336	0.3133	0.5540	0.7777
fub	FIUbasePL2c9	0.3199	0.3738	0.3601	0.6120	0.7351	0.2659	0.3206	0.2915	0.5020	0.6862
UTD_SLP_Lab	SplBaseT	0.3077	0.3688	0.3706	0.5960	0.7152	0.2473	0.3195	0.3012	0.4760	0.6569
KobeU-Seki	ku	0.3035	0.3602	0.3531	0.5820	0.7053	0.2475	0.3051	0.2806	0.4960	0.6585
KU	kunlpKLtt	0.2791	0.3568	0.3487	0.5700	0.7784	0.2263	0.3042	0.2815	0.4520	0.6955
USI	run0	0.2567	0.3363	0.3289	0.4020	0.5472	0.2048	0.2604	0.2523	0.3060	0.4605
feup_irlab	feupB	0.2518	0.3190	0.3243	0.5800	0.7133	0.2006	0.2660	0.2573	0.4360	0.5745
york	york08bb2	0.2074	0.2923	0.2863	0.5540	0.7954	0.1700	0.2489	0.2343	0.4520	0.7308

Table 7: Baseline task: Topic-Relevance and Opinion-Finding - Title only - using the TREC 2008 new topics. All runs are automatic.

documents, categorised in the user display[2]. Evaluation can then be carried out in a more straightforward adhoc document-ranking manner (e.g., using MAP).

3.1 Topics and Assessment

The polarity task shared the same topics as the opinion-finding task. The participating groups were also asked to use all 150 topics: the 50 new topics, as well as the 100 queries from the TREC 2006 and TREC 2007 opinion-finding tasks. In particular, for each topic, a participating system should retrieve and rank all the positive opinionated documents. Then, for each topic, the system should retrieve and rank all the negative opinionated documents. To minimise the number of submitted run files, the groups were asked to concatenate the two runs together in one run file, separated by a blank line. We also required that mixed opinionated documents, i.e. documents containing both positive and negative opinions, should not be listed in the positive (resp. negative) rankings of retrieved documents. The polarity runs were assessed using the same pool described in Section 2.2.

3.2 Results and Main Approaches

In a similar vein to the opinion-finding task, the groups were permitted to submit up to 2 runs to the polarity task, using their own previously submitted baseline runs. A compulsory automatic title-only run was required. In addition, they could submit up to 2 runs using each of the five provided standard topic-relevance baselines (see Table 1). As a consequence, each group could submit up to 12 polarity runs. TREC received a total of 87 polarity runs from 16 groups. Of these runs, 59 used one of the five standard provided baselines, 5 runs had N/A for the baseline (i.e. again, their system did not necessarily separate topic-relevance from polarity detection), and the remaining 23 runs used a baseline run from the corresponding group. For the 59 runs using one of the standard baselines, Table 14 shows the breakdown of runs per standard baseline. Similar to the opinion-finding task, baseline4, which has the highest topic-relevance effectiveness among the provided five standard baselines, was the most frequently used provided baseline. All the submitted runs were automatic runs. For the provided topics, each submitted run consisted of the top 1000 retrieved positive opinionated permalink documents for each topic, followed by the top 1000 retrieved negative opinionated permalink documents for each topic.

Baseline	Number of submitted runs
baseline1	10
baseline2	11
baseline3	11
baseline4	16
baseline5	11
Total	59

Table 14: Breakdown of the submitted polarity runs using one of the provided five standard baselines.

First, we assessed the effectiveness of the 41 submitted baselines in finding positive (pos) and negative (neg) polarised opinions. Moreover, to have an overall retrieval performance for each run, we compute the mean of the positive and negative measures for each run, denoted Mix (e.g. Mix MAP). Table 15 provides

[2]The Opinmind.com search engine used to do this.

			Topic-Relevance					Opinion-Finding				
Group	Run	Fields	MAP	R-prec	bPref	P@10	MRR	MAP	R-prec	bPref	P@10	MRR
KLE	KLEPsgFeedT	T	0.4954	0.5150	0.5364	0.7920	0.9058	0.4052	0.4366	0.4314	0.6440	0.8184
UAms_De_Rijke	uams08n1o1	T	0.4644	0.4867	0.5034	0.7620	0.8892	0.3797	0.4176	0.4117	0.6620	0.8052
UIC_IR_Group	uicirnoa	T	0.4403	0.4804	0.5062	0.7700	0.8667	0.3438	0.3956	0.3929	0.5880	0.7480
UniNE	UniNEBlog1	TD	0.4344	0.4608	0.4662	0.6440	0.8199	0.3565	0.3887	0.3677	0.5540	0.7605
UoGtr	uogBLProxCE	T	0.4219	0.4548	0.4481	0.7060	0.8228	0.3531	0.3840	0.3646	0.6100	0.7723
THUIR	THUrelTwpmf	T	0.4067	0.4565	0.4625	0.6940	0.8263	0.3313	0.3942	0.3749	0.5900	0.7487
BUPT_pris_	prisba	T	0.4065	0.4506	0.4561	0.6780	0.8290	0.3346	0.3876	0.3684	0.5580	0.7456
DUTIR	DUT08BRun1	T	0.3617	0.4188	0.4345	0.6540	0.7633	0.2974	0.3586	0.3598	0.5420	0.7204
iitkgp	IITKGPNOSPAM	T	0.3598	0.4090	0.4394	0.7400	0.8817	0.2988	0.3664	0.3642	0.5720	0.7955
IU-SLIS	wdoqsBase	T	0.3431	0.3918	0.4001	0.7280	0.8636	0.2818	0.3367	0.3215	0.5900	0.7551
UWaterlooEng	UWBase2	T	0.3309	0.3824	0.3875	0.6380	0.8127	0.2753	0.3391	0.3249	0.5160	0.7254
aic-dcu	DCUCDVPtbl	T	0.3303	0.3671	0.3601	0.6520	0.7783	0.2875	0.3280	0.3089	0.5560	0.7066
UTD_SLP_Lab	SplBaseTD	TD	0.3298	0.3751	0.3787	0.6380	0.7423	0.2682	0.3305	0.3133	0.5200	0.6618
UIUC	UIUCb08uwTtl	T	0.3240	0.3766	0.3771	0.6800	0.8223	0.2723	0.3336	0.3133	0.5540	0.7777
fub	FIUbasePL2c9	T	0.3199	0.3738	0.3601	0.6120	0.7351	0.2659	0.3206	0.2915	0.5020	0.6862
KobeU-Seki	ku	T	0.3035	0.3602	0.3531	0.5820	0.7053	0.2475	0.3051	0.2806	0.4960	0.6585
KU	kunlpKLtt	T	0.2791	0.3568	0.3487	0.5700	0.7784	0.2263	0.3042	0.2815	0.4520	0.6955
USI	run0	T	0.2567	0.3363	0.3289	0.4020	0.5472	0.2048	0.2604	0.2523	0.3060	0.4605
feup_irlab	feupB	T	0.2518	0.3190	0.3243	0.5800	0.7133	0.2006	0.2660	0.2573	0.4360	0.5745
york	york08bb2	T	0.2074	0.2923	0.2863	0.5540	0.7954	0.1700	0.2489	0.2343	0.4520	0.7308

Table 8: Baseline task: Topic-Relevance and Opinion-Finding using the TREC 2008 new topics. All runs are automatic.

Group	Run	Fields	Baseline	MAP	Δ MAP	R-prec	bPref	P@10	MRR
KLE	KLEDocOpinT	T	N/A	0.4569	N/A	0.4797	0.4791	0.7200	0.8503
IU-SLIS	top3dt1mRd	T	N/A	0.4335	N/A	0.4618	0.4428	0.6780	0.8483
aic-dcu	DCUCDVPgoo	TD	baseline4	0.4155	8.71%	0.4479	0.4411	0.6800	0.8218
UIC_IR_Group	uicop2bl4r	T	baseline4	0.4067	6.41%	0.4527	0.4338	0.6160	0.7528
fub	FIUBL4DFR	T	baseline4	0.4006	4.81%	0.4447	0.4281	0.6240	0.8097
UoGtr	uogOP4intL	T	baseline4	0.3964	3.72%	0.4370	0.4236	0.6400	0.8137
DUTIR	DUTIR08Run4	T	DUT08BRun2	0.3902	31.60%	0.4257	0.4191	0.6620	0.8082
UTD_SLP_Lab	NOpMM47	TD	baseline4	0.3844	0.58%	0.4258	0.4158	0.6300	0.7908
UAms_De_Rijke	uams08n1o1sp	T	uams08n1o1	0.3823	0.68%	0.4204	0.4139	0.6580	0.8052
THUIR	THUopnTmfRmf	T	THUrelTwpmf	0.3522	6.31%	0.4104	0.3902	0.6320	0.7347
UniNE	UniNEopZ1	TD	UniNEBlog1	0.3418	-4.12%	0.3961	0.3661	0.5840	0.7859
UWaterlooEng	UWnb4Op	T	baseline4	0.3381	-11.54%	0.3718	0.3613	0.6060	0.8231
BUPT_pris_	prisoa1	T	prisba	0.3344	-0.06%	0.3868	0.3679	0.5560	0.7539
USI	opin1kl	T	baseline1	0.3122	-3.61%	0.3584	0.3390	0.5460	0.7062
iitkgp	KGPPOS1	TD	IITKGPTITLE1	0.3005	2.39%	0.3735	0.3633	0.6260	0.8024
KobeU-Seki	kuo	T	ku	0.2704	9.25%	0.3259	0.2978	0.5380	0.7058
york	york08bo1a	T	baseline1	0.2600	-19.73%	0.3160	0.3033	0.3960	0.4817
SUNY_Buffalo	UBop1	TD	N/A	0.1872	N/A	0.2184	0.2259	0.3140	0.4051
KU	kunlpKLttOc	T	kunlpKLtt	0.1752	-22.58%	0.2609	0.2386	0.5200	0.7390

Table 9: Opinion-Finding task: Any baseline, any topic fields, using the TREC 2008 new topics. Ranked by (opinion-finding) MAP. N/A denotes a run by a system that cannot separate the topic-relevance and opinion-finding components. All runs are automatic. Δ MAP denotes the percentage increase in opinion-finding MAP that the opinion-finding run achieved over the opinion-finding MAP of the corresponding baseline run.

Group	Run	Fields	Baseline	MAP	Δ MAP	R-prec	bPref	P@10	MRR
DUTIR	DUTIR08Run4	T	DUT08BRun2	0.3902	31.60%	0.4257	0.4191	0.6620	0.8082
UAms_De_Rijke	uams08n1o1sp	T	uams08n1o1	0.3823	0.68%	0.4204	0.4139	0.6580	0.8052
UoGtr	uogOPb2ofL	T	uogBLProxCE	0.3709	5.04%	0.4049	0.3824	0.6380	0.8114
THUIR	THUopnTmfRmf	T	THUrelTwpmf	0.3522	6.31%	0.4104	0.3902	0.6320	0.7347
UniNE	UniNEopZ1	TD	UniNEBlog1	0.3418	-4.12%	0.3961	0.3661	0.5840	0.7859
BUPT_pris_	prisoa1	T	prisba	0.3344	-0.06%	0.3868	0.3679	0.5560	0.7539
aic-dcu	DCUCDVPtol	T	DCUCDVPtbl	0.3299	14.75%	0.3679	0.3553	0.6360	0.7689
IU-SLIS	wdqfdt1mRd	TDN	wdoqlnvN	0.3127	13.38%	0.3702	0.3518	0.6200	0.8035
iitkgp	KGPPOS1	TD	IITKGPTITLE1	0.3005	2.39%	0.3735	0.3633	0.6260	0.8024
fub	FIUPL2c9DFR	T	FIUbasePL2c9	0.2951	10.98%	0.3507	0.3161	0.5640	0.7288
UWaterlooEng	UWopinion2	T	UWBase2	0.2892	5.05%	0.3361	0.3222	0.5840	0.7832
KobeU-Seki	kuo	T	ku	0.2704	9.25%	0.3259	0.2978	0.5380	0.7058
KU	kunlpKLttOc	T	kunlpKLtt	0.1752	-22.58%	0.2609	0.2386	0.5200	0.7390
USI	opin0kl	T	run0	0.1484	-27.54%	0.1868	0.1736	0.2660	0.3757

Table 10: Opinion-Finding task: Own baseline, any topic fields, using the TREC 2008 new topics. Ranked by MAP. All runs are automatic.

	baseline1	baseline2	baseline3	baseline4	baseline5	improvement	
Baseline	0.3239	0.2639	0.3564	0.3822	0.2988	mean	stdev
TREC median	0.3493	0.2705	0.3705	0.3846	0.3010	+0.76%	0.73%

Table 11: Median opinion MAP over each of the 5 standard baselines and median average improvement for the TREC 2008 topics.

Group	Run	Fields	MAP	Δ MAP
KLE	B4PsgOpinAZN	T	0.4189	9.60%
aic-dcu	DCUCDVPgoo	TD	0.4155	8.71%
UIC_IR_Group	uicop2bl4r	T	0.4067	6.41%
IU-SLIS	b4dt1mRd	T	0.4023	5.26%
fub	FIUBL4DFR	T	0.4006	4.81%
UoGtr	uogOP4intL	T	0.3964	3.72%
UTD_SLP_Lab	NOpMM47	TD	0.3844	0.58%
UWaterlooEng	UWnb4Op	T	0.3381	-11.54%
iitkgp	KGPBASE4	T	0.2852	-25.38%
UAms_De_Rijke	uams08b4pr	T	0.1369	-64.18%
UniNE	UniNEopLRb4		0.2341	-38.75%

Table 12: Opinion-Finding task: Results for runs using standard baseline4, which has the highest topic-relevance and opinion-finding MAP. Ranked by Δ MAP, using the TREC 2008 new topics. No topic fields were specified for run UniNEopLRb4. All runs are automatic.

Group	Approach of	Fields	MAP		Δ MAP	
			Mean	σ	Mean	σ
UIC_IR_Group	uicop1bl1r	T	0.3614	0.04	11.76%	6.93%
KLE	B1PsgOpinAZN	T	0.3565	0.05	9.67%	0.77%
UoGtr	uogOP1PrintL	T	0.3412	0.04	5.21%	5.10%
UTD_SLP_Lab	NOpMM107	TD	0.3273	0.04	0.76%	0.73%
UWaterlooEng	UWnb1Op	T	0.3215	0.02	-0.14%	7.86%
fub	FIUBL1DFR	T	0.2938	0.13	-11.16%	35.62%
UniNE	UniNEopLRb1	T	0.2118	0.02	-34.60%	2.31%
UAms_De_Rijke	uams08b1pr	T	0.1378	0.03	-57.41%	8.02%

Table 13: Opinion-Finding task: Results for runs using all 5 standard baselines, ranked by Mean Δ MAP, using the TREC 2008 new topics. σ denotes the standard deviation. All runs are automatic.

the average best, median, and worst MAP and P@10 measures for each topic, across all submitted 41 baseline runs. In particular, we observed that the retrieval performance of the systems on the TREC 2007 topics was markedly higher than those obtained on the TREC 2006 and TREC 2008 topics when searching for positive opinionated documents, using both MAP and P@10. In contrast, the retrieval performance of the systems on the TREC 2008 topics was higher than those obtained on the TREC 2006 and TREC 2007 topics when searching for negative opinionated documents using both MAP and P@10 evaluation measures. Overall, there is no clear evidence that the three topic sets have different difficulty levels. Table 16 provides the average best, median, and worst MAP and P@10 measures for each topic across all 87 submitted polarity runs. The TREC 2007 topic set appeared to be the easiest for the retrieval of positive opinionated documents, while the three topic sets showed the same level of difficulty when searching for negative opinionated documents. Table 17 provides the average best, median, and worst MAP and P@10 measures for each topic, across all 2006-2008 years (150 topics), for all submitted 41 baseline and 87 polarity runs.

Similar to the opinion-finding task, to avoid any bias towards old topics, in the following, we focus on the performances of the submitted 87 polarity runs on the 50 new TREC 2008 unseen queries. Using MAP, each run is evaluated in terms of its ability to rank positive (resp. negative) opinionated permalinks higher up in the ranking. In order to have an overall performance for each run, we compute the mean of the positive and negative MAPs of each run (denoted Mix MAP), and rank them accordingly. Regardless of the used baseline and the query type, Table 18 shows the best-scoring polarity run for each group in terms of the mean of the positive and negative opinion-finding MAPs of each run (i.e. Mix MAP), sorted in decreasing order. The P@10 measure is also reported. When applicable, the table also compares the Mix MAP of the run to the Mix MAP achieved by its underlying topic-relevance baseline (denoted Mix Δ MAP in the table). A relative increase in performance indicates that the used polarity detection features were useful. However, in most cases, we observe a relative decrease in performance, suggesting that most of the deployed polarity techniques by the participating groups were not successful. Actually, this is also apparent from Tables 15 and 16 where, on average, the submitted baseline systems achieved a higher polarity effectiveness than the submitted polarity runs. Table 19 shows the best-scoring polarity run for each group in terms of Mix MAP, when the group used one of its own submitted baseline runs, regardless of the query type. We observe the same trends, namely that most of the participating systems did not improve the polarity finding effectiveness of the underlying baselines. Overall, we conclude that similar to TREC 2007 [2], the polarity search task appears to be a challenge to most participating systems.

Similar to the analysis performed in Section 2.3, to see the most effective and stable polarity opinion detection techniques, we investigated the extent to which a given polarity opinion finding technique improved the polarity finding MAP of all the five provided standard baselines. Overall, 10 sets of runs using all five standard baselines were submitted by 8 groups. Table 20 shows the median of their improvements over each standard baseline. Table 21 shows the best polarity approach from each of the 8 groups, ranked by the mean of their relative improvements over the five standard baselines, taking into account both their positive and negative polarity opinion retrieval (see column Mean Mix Δ MAP). Only the approach by the KLE group had on average improved the polarity performance of the five provided runs, followed by the approach by the UoGtr group, albeit to some less extent. Both groups used a straightforward extension to their opinion-finding approaches. Indeed, similar to opinion-finding retrieval, the KLE system calculated a positive/negative score of a blog post using the Amazon Review data, while the UoGtr group extended their dictionary-based approach to weight terms according to their positive (resp. negative) opinionated discriminability.

Finally, for the 87 submitted polarity runs, we computed the correlation between the polarity MAP and the topic-relevance MAP. Since for each polarity run there is a positive or a negative part, we correlated using the appropriate run's part (e.g. we correlated negative AP, calculated on the negative MAP run with topic relevance AP, calculated on the negative part of the run). In terms of finding positive opinionated blog posts, the overall rankings of systems are very similar (Spearman's ρ=0.9144 and Kendall's τ=0.7856). A similar high correlation is also observed for negative opinion-finding (Spearman's ρ=0.9341 and Kendall's τ=0.7909). This suggests that the effectiveness of polarity retrieval is strongly dependent on the topic-relevance effectiveness.

4. BLOG DISTILLATION TASK

4.1 Task and Topics

The blog distillation task was first introduced in TREC 2007 [2]. It addresses a search scenario where the user aims to find a blog to follow or read in their RSS reader. This blog should be *principally devoted* to topic X over the timespan of the collection. For example, Google's RSS reader provides an integrated blog search tool to allow users to easily find new blogs of interest. Unlike the blog post search tasks, the blog distillation task aims to rank blogs (aggregates of blog posts) instead of permalink documents.

Like in TREC 2007, the topics were contributed and judged by the participating groups. However, the topic creation guidelines given to the participating groups have been tightened up. Indeed, based on experience from TREC 2007, the participating groups were explicitly asked to avoid topics that are too general, with too many relevant documents (e.g. Linux), or topics with temporal aspects, i.e. topics likely to be of interest only in a specific period of time (Christmas). Each participating group was asked to contribute 6 topics, along with some relevant blogs. Similar to TREC 2007, to help the participating groups in creating their topics, the organisers have provided a standard search system for documents on the Blogs06 collection using the Terrier search engine [4], which also displays the blogs for each document, as well as all the documents for a given blog. Overall, 11 participating groups sent a total of 66 topics. From each group, TREC selected 4 or 5 topics to form a set of 50 new topics.

4.2 Assessments

Relevance judgements were conducted by 11 participating groups, using a slight improvement of the TREC Blog track 2007's community judgements system interface [2, 5]. In particular, the assessors were asked to mark splogs (spam blogs), and to differentiate between relevant and highly relevant blogs. This allows the use of measures such as nDCG, and to have a better analysis of the blog distillation task's relevance assessments. As a consequence, the guidelines instructed to the assessors of each participating group were to read the query and its narrative, and to judge each blog in the provided pool. Relevance judgements were made on a four-point scale:

Spam: This is a spam blog (splog).

Not relevant: I would definitely not subscribe to this feed.

	2006 (851-900)			2007 (901-950)			2008 (1001-1050)		
	MAP_{pos}	MAP_{neg}	MAP_{mix}	MAP_{pos}	MAP_{neg}	MAP_{mix}	MAP_{pos}	MAP_{neg}	MAP_{mix}
median	0.0733	0.0618	0.0690	0.1627	0.0605	0.1137	0.1070	0.1031	0.1064
best	0.1589	0.1556	0.1429	0.3024	0.1476	0.2063	0.2487	0.1996	0.2070
worst	0.0062	0.0053	0.0067	0.0220	0.0032	0.0143	0.0149	0.0082	0.0133
	$P@10_{pos}$	$P@10_{neg}$	$P@10_{mix}$	$P@10_{pos}$	$P@10_{neg}$	$P@10_{mix}$	$P@10_{pos}$	$P@10_{neg}$	$P@10_{mix}$
median	0.1220	0.0820	0.1110	0.2200	0.0580	0.1490	0.1520	0.1380	0.1550
best	0.3480	0.2860	0.2610	0.4760	0.2520	0.3120	0.3980	0.3140	0.3040
worst	0.0000	0.0020	0.0030	0.0120	0.0000	0.0080	0.0080	0.0080	0.0140

Table 15: Baseline runs: Best, median and worst positive, negative, and mixed MAP and P@10 measures of the 2008 participating groups across the three topic sets.

	2006 (851-900)			2007 (901-950)			2008 (1001-1050)		
	MAP_{pos}	MAP_{neg}	MAP_{mix}	MAP_{pos}	MAP_{neg}	MAP_{mix}	MAP_{pos}	MAP_{neg}	MAP_{mix}
median	0.0796	0.0638	0.0751	0.1734	0.0628	0.1241	0.0899	0.0678	0.0808
best	0.3890	0.5178	0.4019	0.5405	0.5332	0.4763	0.2723	0.2365	0.2297
worst	0.0033	0.0013	0.0033	0.0025	0.0007	0.0030	0.0027	0.0014	0.0030
	$P@10_{pos}$	$P@10_{neg}$	$P@10_{mix}$	$P@10_{pos}$	$P@10_{neg}$	$P@10_{mix}$	$P@10_{pos}$	$P@10_{neg}$	$P@10_{mix}$
median	0.1640	0.1260	0.1530	0.2740	0.1120	0.2000	0.1640	0.1300	0.1550
best	0.7120	0.7460	0.6190	0.7940	0.6200	0.6160	0.4940	0.4060	0.4030
worst	0.0000	0.0020	0.0010	0.0000	0.0000	0.0010	0.0020	0.0000	0.0020

Table 16: Polarity runs: Best, median and worst positive, negative, and mixed MAP and P@10 measures of the 2008 participating groups across the three topic sets.

Relevance Scale	Level	Nbr. of Documents	Avg.
Highly Relevant	2	792	15.84
Relevant	1	1151	23.02
Not Relevant	0	13979	279.58
Spam	-1	2080	41.6
Total	-	18002	360.04

Table 22: Blog distillation task: Distribution of relevance levels in the pool.

Relevant: This contains enough on-topic posts such that I would probably subscribe to it in my RSS reader.

Highly relevant: I would definitely subscribe to this blog for that topic.

4.3 Results

Participants were allowed to submit up to 4 runs, including a compulsory automatic title-only run. Each run had blogs ranked by their likelihood of having a principal (recurring) interest in the topic. Given the number of blogs in the collection (just over 100k blogs), each run consisted of up to 100 returned blogs for each topic. Overall, 43 runs were submitted by 12 participating groups [3]. All of the submitted runs were automatic. A pool was then formed by NIST including the top 50 documents from two runs per participant. Table 22 shows the distribution of relevance levels across all topics. On average, each topic had about 16 highly relevant blogs, which are principally devoted to the topic of the query.

Figure 3 shows the distribution of the number of blogs in the pool for the different relevance levels per topic. The topics are ordered by the descending sum of their corresponding relevant and highly relevant documents. The general topics appear early in the graph, while those with very few relevant documents appear late on the X axis. Considering the presence of spam in the pool, we note some variance in the number of returned splogs across the 50 used topics. For example, some topics had more than 118 spam blogs in the pool (e.g. "subprime lending" (1058), "celebrity babies" (1078), or "3d cities globes" (1086)), while others had very little corresponding spam in the pool (e.g. "road cycling" (1077), "jazz" (1064), or "Hubei" (1095)). We also note that there appears to be some variance in the number of relevant blogs across the 50 used topics. Indeed, some topics had very little relevant blogs in the pool (e.g. "beach volleyball" (1062) had only two relevant blogs and no highly relevant ones), while others had a high number of relevant and highly relevant blogs (e.g. topic "Firefox" (1059), which had 40 relevant blogs and 116 highly relevant ones). Other topics that had over 100 relevant and highly relevant blogs are topics "cooking recipes" (1053) and "SEO" (1060). Such large numbers of relevant blogs were observed even after tightening up the topics creation guidelines provided to the participating groups, so as to precisely avoid such a situation.

[3]One group who participated in the topics creation and assessments, UCSC, did not submit runs, while two groups submitted runs but did not participate in the topics creation and relevance assessments phases, namely, KLE and IITKGP.

	nDCG	MAP	MRR
Best	0.6600	0.4379	0.9583
Median	0.4492	0.2239	0.7213

Table 23: Best, median and worst nDCG, MAP & MRR measures for the 43 submitted runs to the blog distillation task.

The blog distillation task is a precision-oriented search task where systems that retrieve the highly relevant documents should be favoured. Therefore, in evaluating the runs, we report the nDCG evaluation measure, which takes into account the graded relevance levels. We also report the classical retrieval measures such as MAP and precision at fixed ranks. Table 23 provides the average best and median nDCG, MAP, and MRR measures for each topic, across all 43 submitted runs.

Table 24 shows the best-scoring automatic title-only run from each participating group in terms of nDCG, and sorted in decreasing order. MAP(2) denotes the MAP of the run, when only the judged highly relevant blogs are considered to be relevant. Table 25 shows the best run from each group, regardless of the topic length used. Note that most of the 43 submitted runs were title-only runs.

	Baseline runs			Polarity runs		
	MAP_{pos}	MAP_{neg}	MAP_{mix}	MAP_{pos}	MAP_{neg}	MAP_{mix}
median	0.1143	0.0751	0.0964	0.1143	0.0648	0.0933
best	0.2367	0.1676	0.1854	0.4006	0.4292	0.3693
worst	0.0144	0.0055	0.0114	0.0028	0.0011	0.0031
	$P@10_{pos}$	$P@10_{neg}$	$P@10_{mix}$	$P@10_{pos}$	$P@10_{neg}$	$P@10_{mix}$
median	0.1647	0.0927	0.1383	0.2007	0.1227	0.1693
best	0.4073	0.2840	0.2923	0.6667	0.5907	0.5460
worst	0.0067	0.0033	0.0083	0.0007	0.0007	0.0013

Table 17: Best, median, worst of baseline and polarity runs over all 150 topics.

Group	Run	Fields	Baseline	Mix			Positive			Negative		
				MAP	Δ MAP	P@10	MAP	Δ MAP	P@10	MAP	Δ MAP	P@10
IU-SLIS	top3dt1mP5	T	N/A	0.1677	N/A	0.2170	0.1752	N/A	0.2040	0.1601	N/A	0.2300
KLE	KLEPolarity	T	N/A	0.1662	N/A	0.2020	0.1828	N/A	0.2360	0.1496	N/A	0.1680
aic-dcu	DCUCDVPgpo	TD	baseline4	0.1547	9.70%	0.1900	0.1612	5.22%	0.2000	0.1483	15.14%	0.1800
KobeU-Seki	kup4	T	baseline4	0.1448	2.68%	0.1820	0.1566	2.22%	0.1980	0.1329	3.18%	0.1660
THUIR	THUpolTmfPNR	T	THUrelTwpmf	0.1353	7.16%	0.1870	0.1289	6.27%	0.1880	0.1417	7.92%	0.1860
UoGtr	uogPL41	T	baseline4	0.1348	-4.41%	0.1640	0.1394	-9.01%	0.1700	0.1301	1.01%	0.1580
UWaterlooEng	UWnb1Pol	T	baseline1	0.1278	0.71%	0.1780	0.1430	4.84%	0.2040	0.1126	-4.17%	0.1520
iitkgp	KGPPOL1	T	IITKGPTITLE1	0.1139	-6.15%	0.1990	0.1304	-1.95%	0.2300	0.0975	-11.12%	0.1680
UTD_SLP_Lab	NTrMM47P	TD	baseline4	0.1129	-19.94%	0.2130	0.1323	-13.64%	0.2220	0.0934	-27.48%	0.2040
UIC_IR_Group	uicpolrun1	T	N/A	0.1099	N/A	0.2400	0.1594	N/A	0.3000	0.0604	N/A	0.1800
UniNE	UniNEpolLR1	TD	UniNEBlog1	0.0775	-41.33%	0.1780	0.0882	-35.90%	0.2000	0.0667	-47.31%	0.1560
fub	FIUpBL3DFR	T	baseline3	0.0723	-45.26%	0.1610	0.0618	-55.09%	0.1760	0.0828	-34.60%	0.1460
SUNY_Buffalo	UBpol1	T	N/A	0.0661	N/A	0.0820	0.0752	N/A	0.1080	0.0570	N/A	0.0560
tno	tnobase1	D	baseline1	0.0449	-64.62%	0.0990	0.0544	-60.12%	0.1360	0.0353	-69.96%	0.0620
KU	kunlpKLttPs	T	kunlpKLtt	0.0416	-54.39%	0.1560	0.0542	-38.34%	0.1900	0.0291	-69.21%	0.1220
DUTIR	DUTIR08Run2P	T	DUT08BRun2	0.0301	-73.43%	0.1500	0.0352	-72.28%	0.1840	0.0250	-74.87%	0.1160

Table 18: Polarity task: Any baseline, any topic fields, using the TREC 2008 new topics. Ranked by Mix MAP.

Indeed, there were 36 submitted runs using the title-only field, 3 submitted runs using the title, description and narrative fields, and 4 submitted runs using the title and description fields. However, Table 25 shows that 3 out of the top 5 runs used more than the title field of the topics.

The overall rankings of systems using either the nDCG or the MAP measures were very similar. Indeed, we observed a very high Spearman's correlation of $\rho = 0.9807$ for the 43 submitted runs (Kendall's τ distance leads to a similar high correlation of $\tau = 0.8936$). If only the highly relevant documents are considered in ranking the systems (i.e. systems are ranked by MAP(2)), then the ranking of systems is very similar to the one obtained using both relevant and highly relevant documents (i.e. using MAP measure): $\rho = 0.9461, \tau = 0.7984$ for the 43 submitted runs. This suggests that the ranking of systems are almost identical whether using nDCG, MAP or MAP(2) [7].

4.4 Participants Approaches

Almost all groups indexed only the permalinks component of the Blogs06 collection. The only exceptions are the CMU and DUTIR groups who indexed both blogs and permalinks components, and the WHU group who experimented with two indexes: one based on feeds only and one based on the permalinks component only.

In terms of retrieval approaches, we noted an interesting trend, namely the use of expert search techniques to rank blogs. The idea, first proposed by the University of Glasgow in TREC 2007 [8], was used by three groups in TREC 2008: UAms, UoGtr and USI. Both UoGtr and USI use the Voting Model to rank blogs [8]. Using an expert search approach, the UAms group explored the use of various external corpora to improve the effectiveness of query expansion. They also used several blog characteristics such as the number of comments, post length, or the posting time to estimate the strength of association between a post and a blog. Starting from their Voting Model for blog search, the UoGtr group added a component with a focus on a balanced and neutral retrieval that does not favour prolific bloggers. They also investigated the use of a feature which ascertains if the retrieved posts in a given blog for a topic are spread across the timespan of the Blogs06 collection. The idea is to model the notion of recurring interests. Finally, to further enrich the topics, UoGtr employed a collection enrichment technique, using the Wikipedia corpora. They observed that while each of their deployed techniques improved the effectiveness of their baseline run, the latter had an only average retrieval effectiveness. Finally, on top of an expert search approach used as a baseline, the USI group tested the use of structure-based evidence besides content in a Rank Learning approach. However, they observed that the Rank Learning model appears to be very sensitive to the properties of the data set, and did not perform well in their experiments.

Other retrieval groups, such as WHU, tested whether using folksonomies to expand the queries improves the retrieval effectiveness. They showed that the approach is only beneficial with a Feeds-based index, while it is detrimental to retrieval when a Permalinks-based index is used. The FEUP group investigated two features based on temporal evidence – temporal span and temporal dispersion. The temporal span of a topic in a blog corresponds to the period between the newest relevant post and the oldest relevant post. Both features were combined with a baseline BM25 run based on Terrier. Finally, the UMass group used a query likelihood language modelling approach. Recent posts are boosted higher in the aggregation of the scores of relevant posts.

Various groups implemented their solutions on top of existing information retrieval platforms such as Lucene (IITKGP), Indri/Lemur (CMU, Umass), and Terrier (UoGtr, USI, FEUP, WHU), using various document ranking models ranging from BM25 to language modelling, through Divergence From Randomness models. In the following, we provide a detailed description of the methods used by the two top performing groups in the blog distillation task.

Group	Run	Fields	Baseline	Mix			Positive			Negative		
				MAP	Δ MAP	P@10	MAP	Δ MAP	P@10	MAP	Δ MAP	P@10
THUIR	THUpolTmfPNR	T	THUrelTwpmf	0.1353	7.16%	0.1870	0.1289	6.27%	0.1880	0.1417	7.92%	0.1860
UoGtr	uogPLb21	T	uogBLProxCE	0.1274	-0.62%	0.1560	0.1372	-1.15%	0.1700	0.1176	0.00%	0.1420
IU-SLIS	wdqbdt1mP5	T	wdoqsBase	0.1143	9.51%	0.2160	0.1147	6.80%	0.2280	0.1138	12.34%	0.2040
iitkgp	KGPPOL1	T	IITKGPTITLE1	0.1139	-6.15%	0.1990	0.1304	-1.95%	0.2300	0.0975	-11.12%	0.1680
aic-dcu	DCUCDVPtpl	T	DCUCDVPtbl	0.1092	9.88%	0.1550	0.1087	13.35%	0.1380	0.1097	6.61%	0.1720
UWaterlooEng	UWpolarity2	T	UWBase2	0.1078	-0.27%	0.1670	0.1215	9.66%	0.2000	0.0942	-10.63%	0.1340
KobeU-Seki	kup	T	ku	0.0994	9.83%	0.1650	0.1056	13.79%	0.1740	0.0933	5.78%	0.1560
UniNE	UniNEpolLR1	TD	UniNEBlog1	0.0775	-41.33%	0.1780	0.0882	-35.90%	0.2000	0.0667	-47.31%	0.1560
fub	FIUpPL2DFR	T	FIUbasePL2c9	0.0506	-46.91%	0.1290	0.0529	-48.94%	0.1680	0.0483	-44.55%	0.0900
KU	kunlpKLttPs	T	kunlpKLtt	0.0416	-54.39%	0.1560	0.0542	-38.34%	0.1900	0.0291	-69.21%	0.1220
DUTIR	DUTIR08Run2P	T	DUT08BRun2	0.0301	-73.43%	0.1500	0.0352	-72.28%	0.1840	0.0250	-74.87%	0.1160

Table 19: Polarity task: Own baseline, any topic fields, using the TREC 2008 new topics. Ranked by Mix MAP.

negative	baseline1	baseline2	baseline3	baseline4	baseline5	improvement	
Baseline	0.1175	0.0865	0.1266	0.1288	0.1085	mean	stdev
TREC median	0.0597	0.0457	0.0743	0.0677	0.0453	-48.49%	2.66%
positive	baseline1	baseline2	baseline3	baseline4	baseline5	improvement	
Baseline	0.1364	0.0951	0.1376	0.1532	0.1229	mean	stdev
TREC median	0.0953	0.0547	0.0955	0.0973	0.0708	-36.79%	17.48%

Table 20: Median negative and positive MAP over each of the 5 standard baselines and median average improvement for the TREC 2008 topics.

The KLE group used two scores for a given blog. The first score is the average score of all posts in the blog. The KLE system assumes that the blog that has many relevant posts is more relevant. The second score is the average score of the top N posts that have the highest relevance scores. The KLE system assumes that the top N posts best represent the topic of the blog. The topic-relevance score of each post is calculated using a language modeling approach. To estimate the query model, KLE used the top M blogs in the feedback step. This method increases the diversity of feedback documents, and results in a more effective query model.

The CMU group explored document representation, retrieval models, query expansion and spam filtering. CMU's retrieval system, based on Indri, used a combined index of the permalink and blog documents, distinctly weighting text from various parts of the HTML and XML. Two retrieval models were applied to blog distillation: the large document model, where each blog is viewed as a single document; and the small document model, where a blog is represented as a collection of individual entry documents. Similarly to last year's results, CMU's best performing run used a query expansion method that leverages the link structure in Wikipedia. A spam filtering component was also integrated, which led to further performance improvements.

5. CONCLUSIONS

Back in 2006, when we first proposed the Blog track, our aim was to have a long-term objective for the Blog track, recognising that the richness of the blogosphere and its peculiarities will require several years of investigation before reaching a full understanding of the different blog search tasks, and how they should be effectively addressed. In particular, we proposed to adopt an incremental approach, where we begin with basic blog search tasks and progressively move to more complex search scenarios. We believe that the opinion-finding, its natural polarity extension, as well as the blog distillation tasks are good articulations of real user tasks, albeit basic, in adhoc search behaviour on the blogosphere.

After three years of the Blog track, we believe that we have a good test collection for the opinion-finding task and its polarity extension. In particular, the setting of the TREC 2008 Blog track's opinion-finding and polarity tasks, which provides the participating groups with various standard topic-relevance baselines, on which they can evaluate their opinion-finding techniques, should allow for a better understanding of these tasks and how the opinion-finding performance varies across different baselines. We believe, therefore, that the opinion-finding task in its current form should be discontinued. Instead, we propose to use the notion of opinion as a feature or a dimension of more refined and complex search tasks, as outlined below.

The current blog distillation task only focuses on topical relevance. It does not address the quality aspect of the retrieved blogs. In a position paper, Hearst et al. [9] proposed an interesting refinement of the blog distillation task that takes into account a number of attributes or facets such as the authority of the blog, the trustworthiness of its authors, or the genre of the blog (e.g. opinionated or not) and its style of writing. For example, a user might be interested in blogs to read about a topic X, but where the blogger expresses opinionated viewpoints, backed up by a scientific methodology or evidence. In other words, a user might not be interested in all blogs having a recurring and principal interest in a given topic X, but only those blogs that satisfy a set of criteria or facets.

For TREC 2009, we propose to move to a second phase of the Blog track, where more refined and complex search scenarios will be investigated. In particular, we propose to use a new and larger collection of blogs, Blogs08, which has a much longer timespan than the 11-weeks period covered in the Blogs06 collection. This allows for investigating another important characteristic of the blogosphere, namely the temporal/chronological aspect of blogging, and various related search tasks such as story identification and tracking. One of our proposed tasks for next year is a refinement of the blog distillation task, which addresses the quality aspect through the use of facets.

Acknowledgements

The description of system runs are based on paragraphs contributed by the participating groups. Thanks are also due to the 11 groups who created and assessed this year's blog distillation task topics. Finally, we are grateful to Rodrygo Santos for handling the blog distillation task relevance assessments system and for various editing help with the overview paper.

Group	Approach of	Fields	Mix				Positive				Negative			
			MAP		Δ MAP		MAP		Δ MAP		MAP		Δ MAP	
			Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ
KLE	B1Polarity	T	0.1274	0.02	4.86%	2.69%	0.1370	0.02	6.08%	1.72%	0.1180	0.02	3.51%	7.43%
UoGtr	uogPL11	T	0.1165	0.02	-3.77%	2.28%	0.1226	0.02	-4.62%	2.91%	0.1103	0.01	-2.76%	3.50%
UWaterlooEng	UWnb1Pol	T	0.1119	0.01	-6.70%	8.80%	0.1252	0.01	-1.69%	10.04%	0.0987	0.01	-12.33%	7.71%
UIC_IR_Group	uicpol1bl1	T	0.0941	0.01	-22.10%	8.35%	0.1313	0.02	2.12%	9.83%	0.0568	0.01	-49.60%	7.63%
UTD_SLP_Lab	NTrMM17P	TD	0.0934	0.01	-22.96%	2.53%	0.1068	0.02	-17.51%	4.06%	0.0799	0.01	-29.23%	5.01%
fub	FIUpBL1DFR	T	0.0545	0.02	-55.26%	15.80%	0.0521	0.02	-59.81%	15.01%	0.0569	0.02	-50.18%	17.11%
tno	tnobase1	D	0.0286	0.01	-76.42%	6.12%	0.0312	0.01	-75.93%	8.03%	0.0260	0.01	-77.02%	4.05%
UniNE	UniNEpolLRb1		0.0680	0.01	-43.68%	3.13%	0.0775	0.01	-39.41%	4.70%	0.0584	0.01	-48.49%	2.66%

Table 21: Polarity task: Results for runs using all 5 of the standard baselines. Ranked by Mean Δ Mix MAP, using the TREC 2008 new topics. No fields were specified for run UniNEpolLRb1.

Group	Run	nDCG	MAP	R-prec	bPref	P@10	MRR	MAP(2)
KLE	KLEDistLMT	0.5324	0.3015	0.3601	0.3580	0.4480	0.7977	0.2935
CMU-LTI-DIR	cmuLDwikiSP	0.5170	0.3056	0.3646	0.3535	0.4340	0.8051	0.2750
UAms_De_Rijke	uams08bl	0.4904	0.2638	0.3137	0.3024	0.4200	0.7294	0.2547
uMass	UMassBlog1	0.4777	0.2520	0.3077	0.2944	0.3880	0.7504	0.2561
UoGtr	uogTrBDfeNWD	0.4758	0.2521	0.3121	0.2932	0.4040	0.7425	0.2452
KobeU-Seki	kudb	0.4712	0.2422	0.2947	0.2903	0.3440	0.7469	0.2398
SUNY_Buffalo	UBDist1	0.4694	0.2410	0.2916	0.2855	0.3720	0.6864	0.2413
USI	BM25LenNorm	0.4663	0.2566	0.3144	0.2882	0.3960	0.7016	0.2282
WHU	PermMeWhu	0.4023	0.1898	0.2591	0.2451	0.3180	0.5554	0.1827
feup_irlab	feupbase	0.3478	0.1413	0.1890	0.1690	0.2560	0.5970	0.1621
iitkgp	FEEDKGP	0.3397	0.1539	0.2146	0.1916	0.2680	0.5119	0.1456
DUTIR	DUTIR08DRun1	0.3370	0.1600	0.2293	0.2054	0.2600	0.4543	0.1272

Table 24: Blog distillation task, best run for each group, title-only topics. Ranked by nDCG.

Group	Run	Topic	nDCG	MAP	R-prec	bPref	P@10	MRR	MAP(2)
KLE	KLEDistFBB	TD	0.5443	0.2994	0.3508	0.3224	0.4560	0.7458	0.2852
CMU-LTI-DIR	cmuLDwikiSP	T	0.5170	0.3056	0.3646	0.3535	0.4340	0.8051	0.2750
uMass	UMassBlog3	TD	0.4969	0.2711	0.3286	0.3117	0.4240	0.7612	0.2772
UAms_De_Rijke	uams08bl	T	0.4904	0.2638	0.3137	0.3024	0.4200	0.7294	0.2547
SUNY_Buffalo	UBDist4	TDN	0.4824	0.2633	0.3160	0.3088	0.3820	0.7293	0.2449
UoGtr	uogTrBDfeNWD	T	0.4758	0.2521	0.3121	0.2932	0.4040	0.7425	0.2452
KobeU-Seki	kudb	T	0.4712	0.2422	0.2947	0.2903	0.3440	0.7469	0.2398
USI	BM25LenNorm	T	0.4663	0.2566	0.3144	0.2882	0.3960	0.7016	0.2282
WHU	PermMeWhu	T	0.4023	0.1898	0.2591	0.2451	0.3180	0.5554	0.1827
iitkgp	FEEDKGP1	TD	0.3613	0.1720	0.2484	0.2129	0.3220	0.5077	0.1826
feup_irlab	feupbase	T	0.3478	0.1413	0.1890	0.1690	0.2560	0.5970	0.1621
DUTIR	DUTIR08DRun4	TDN	0.3394	0.1632	0.2365	0.2059	0.2780	0.4298	0.1359

Table 25: Blog distillation task, best run for each group, any topic fields. Ranked by nDCG.

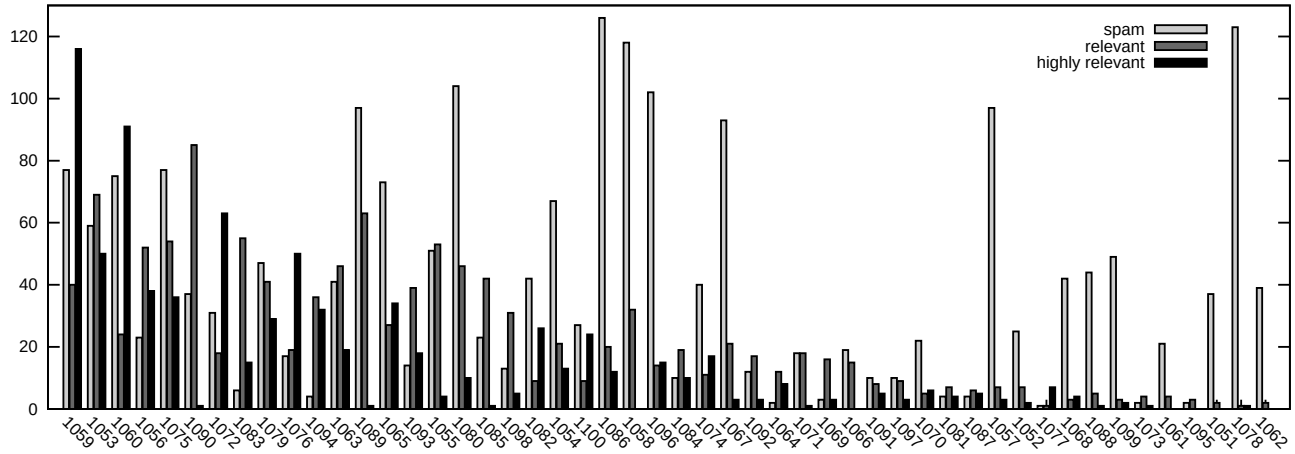

Figure 3: Blog distillation task: Distribution of assessed blogs across all 2008 topics 1051 to 1100 for relevance levels -1 (spam), 1 (relevant), and 2 (highly relevant). Blogs judged as 0 (non-relevant) are omitted for the sake of clarity.

6. REFERENCES

[1] C. Macdonald, I. Ounis, I. Soboroff. Overview of TREC-2007 Blog track. In *Proceedings of TREC-2007*, Gaithersburg, USA, 2008.

[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] 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`

[4] 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.

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

[6] C. Macdonald, B. He, I. Ounis, and I. Soboroff. Limits of opinion-finding baseline systems. In *Proceedings of SIGIR-2008*, Singapore, 2008.

[7] E. Voorhees. Evaluation by highly relevant documents. In *Proceedings of SIGIR-2001*, New Orleans, 2001.

[8] D. Hannah, C. Macdonald,J. Peng,B. He, and I. Ounis. University of Glasgow at TREC2007: Experiments in Blog and Enterprise tracks with Terrier. In *Proceedings of TREC-2007*, Gaithersburg, USA, 2008.

[9] M. Hearst, M. Hurst, and S. Dumais. What Should Blog Search Look Like? In *Proceedings of SSM-2008*, Napa Valley, USA, 2008.