Overview of the TREC-2009 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 Blog track explores the information seeking behaviour in the blogosphere. Thus far, since its inception in 2006 [9], the Blog track addressed two main search tasks based on the analysis of a commercial blog search engine: 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 TREC 2009, the Blog track has been markedly revamped with the use of a new and larger sample of the blogosphere, called Blogs08, which has a 13-month timespan covering a period ranging from 14th January 2008 to 10th February 2009, and the introduction of two new search tasks, addressing more refined and typical search scenarios on the blogosphere:

- *Faceted blog distillation*: A more refined version of the blog distillation task, addressing the quality aspect of the retrieved blogs.
- *Top stories identification*: A task that addresses news-related issues on the blogosphere.

Most of the efforts of the organisers in the Blog track 2009 have been spent on defining the new search tasks, on building a suitable infrastructure to support the investigation of the introduced search tasks, and on establishing an appropriate methodology to evaluate the effectiveness of the submitted runs. The remainder of this paper is structured as follows. Section 2 describes the newly created Blogs08 collection. Section 3 describes the new faceted blog distillation task, and discusses the main obtained results by the participating groups. Section 4 describes the top stories identification task, and summarises the results of the runs and the main effective approaches deployed by the participating groups. Concluding remarks are provided in Section 5.

2. BLOGS08 COLLECTION

Previous incarnations of the TREC Blog track, 2006-2008, used a specially created test collection called Blogs06. It was an 11 week snapshot of 100,000 blogs from late 2005 and early 2006. For TREC 2009, the University of Glasgow created a new test collection, called Blogs08, a markedly larger and more up-to-date sample of the blogosphere. The Blogs08 collection has a much longer timespan period than that of the older Blogs06 collection. The new collection provides a better experimental environment for the faceted blog distillation task, offers the possibility to study the temporal/chronological aspect of blogging, as well as the opportunity to tackle related tasks such as filtering and story/event identification and tracking.

Quantity	Blogs06	Blogs08
Number of Unique Blogs	100,649	1,303,520
First Feed Crawl	06/12/2005	14/01/2008
Last Feed Crawl	21/02/2006	10/02/2009
Number of Permalinks	3,215,171	28,488,766
Total Compressed Size	25GB	453GB
Total Uncompressed Size	148GB	2309GB
Feeds (Uncompressed)	38.6GB	808GB
Permalinks (Uncompressed)	88.8GB	1445GB
Homepages (Uncompressed)	20.8GB	56GB

Table 1: Statistics of the Blogs06 and Blogs08 test collections.

For the creation of the Blogs08 collection, we monitored 1 million blogs on a weekly basis from 14th January 2008 to 10th February 2009. This timespan of over 1 year allowed a good sample of the blogosphere to be obtained, and facilitates studying the structure, properties and evolution of the blogosphere, as well as addressing research tasks such as how the blogosphere responds to events as they happen. In particular, the collection covers the full US election cycle. We included a selection of "top blogs" from a blog directory website, as well as blogs included in Blogs06 and mined from various sources, such as blog search engines. While we did not add any particular spam blog feeds to Blogs08, it is highly likely that it does contain some.

Similarly to Blogs06, Blogs08 includes the XML feed every time a blog was checked. If new permalinks were found when checking this feed, the new permalink was downloaded at least two weeks later (to allow comments to be posted on the permalink). Lastly, at the end of the crawl, the homepage of each blog was downloaded once. The final collection was shipped to the Blog track participants by the University of Glasgow[1]. Table 1 shows the statistics of the final Blogs08 collection, along with comparable figures from Blogs06 [6].

3. FACETED BLOG DISTILLATION TASK

The blog distillation task was first introduced in TREC 2007 [5]. Blog search users often wish to identify blogs about a given topic X, which they can subscribe to and read on a regular basis in their RSS reader. For a given topic X, a retrieval system aims to find blogs that are principally devoted to X over the timespan of the collection. An overview of the retrieval techniques used in the TREC Blog track to build such systems can be found in [5, 10]. However, in its TREC 2007 and TREC 2008 incarnations, the blog distilla-

[1] Further information on obtaining the Blogs08 collection can be found at `http://ir.dcs.gla.ac.uk/test_collections/`

tion task only focused on topical relevance. It did not address the "quality" aspect of the retrieved blogs.

A position paper by Hearst et al. [3] describes a blog search engine user interface that supports *exploratory search*, by means of *facets* that allow the filtering of blogs according to various attributes. Suggested facets may include the opinionated nature of a blog, the trustworthiness of its authors, its style of writing, or its genre. We believe that this goes some way to addressing the quality aspect missing from the previous incarnations of this task.

Following this, for TREC 2009, we introduced a refined blog distillation task, which takes into account facets during retrieval. Firstly, some definitions: a facet is a method of restricting the retrieved results. Each facet has one or more *inclinations*, which allow the user to specify the way in which a facet restriction should be applied. For example, a user might be interested in blogs to read about a topic X, but where the blogger is regarded as trusted – in this case, the facet is trustworthiness, and the active inclination is trustworthy. Hence, 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 the set facet inclinations. Indeed, the new faceted blog distillation task can therefore be summarised as "Find me a *good* blog with a principal, recurring interest in X", where the sought quality of the blogs is characterised through the set facet *inclinations*.

3.1 Task Definition and Topics

The new faceted blog distillation task has the following characteristics: (i) it goes beyond topical-relevance (ii) it integrates a quality aspect in the evaluation of the retrieved blogs and (iii) it mimics an exploratory search task. Each topic has facets of interest attached to it. For TREC 2009, we used an initial set of three facets of varying difficulty, which were all assumed to have binary inclinations for operational simplicity. Namely, the three facets used for TREC 2009 were:

Opinionated: Some bloggers may make opinionated comments on the topics of interest, while others report factual information. A user may be interested in blogs, which show prevalence to opinionatedness. For this facet, the inclinations of interest are 'opinionated' vs 'factual' blogs.

Personal: Companies are increasingly using blogging as an activity for public relations purposes. However, a user may not wish to read such mostly marketing or commercial blogs, and may prefer instead to keep up with blogs that appear to be written in personal time without commercial influences. For this facet, the inclinations of interest are 'personal' vs 'official' blogs.

In-depth: Users might be interested to follow bloggers whose posts express in-depth thoughts and analysis on the reported issues, preferring these over bloggers who simply provide quick bites on these topics, without taking the time to analyse the implications of the provided information. For this facet, the inclinations of interest are 'indepth' vs. 'shallow' blogs (in terms of their treatment of the subject).

The main difficulty of the task for participants consists in identifying a set of features that allow the participating systems to score the extent to which a blog satisfies the set facet inclination (e.g. shallow in terms of its treatment of the subject or personal), and rerank the relevant blogs accordingly. We specifically chose to have an 'opinionated' facet so that participating groups could leverage past track work on blog post opinion-finding [5, 9, 10]. It is of note that while the facets were predefined for TREC 2009, possible future incarnations of this task may require systems to automatically select the facets they consider to be interesting for a given query. 50 new topics were created by NIST assessors. During the topic development, one appropriate facet was chosen for each topic. In particular, the facet Opinionated has been associated to 21 topics, the facet Personal has been associated to 10 topics, and the facet In-depth has been associated to 19 topics. An example of a topic associated with the facet Personal is included in Figure 1.

```
<top>
<num> Number: 1105 </num>

<query> parenting </query>

<desc> Description:
I am looking for blogs that provide advice,
counseling, and information on parenting.
</desc>

<facet> personal </facet>

<narr> Narrative:
Relevant blogs include those from parents,
grandparents, or others involved in
parenting, raising, or caring for children.
Blogs can include those provided by health
care providers if the focus is on children.
Blogs that serve primarily as links to
other sites, or that of themselves, market
products related to children and their
caregivers, are not relevant.
</narr>

</top>
```

Figure 1: Blog track 2009, faceted blog distillation task, topic 1105.

3.2 Assessments and Pools

The blog distillation task is a feed search task. Therefore, the retrieval units are the documents from the feeds components of the Blogs08 test collection. For each topic, the participating groups were asked to supply three rankings of 100 blogs each: one with the first inclination of the facet enabled, one with the second inclination of the facet enabled, and one for a baseline ranking with no facet inclination detection whatsoever enabled. The latter, denoted by 'none', is used as a baseline. For example, for the *Personal* facet, the first ranking would have 100 blogs that the system assesses as being 'personal', the second ranking would have 100 blogs which the system assesses as being 'official', while the third ranking would have 100 blogs which the system assesses as being relevant to the topic, without any consideration for the facet.

We used an assessment procedure inspired by the opinion-finding task in TREC 2006-2009 [5, 9, 10]. In particular, the following scale has been used for the assessment of the returned blogs:

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

0 *Not relevant.* The blog and its posts were examined, and does not contain any interest in the target topic area, or refers to

Relevance Level	# Queries	# Blogs
Not Relevant	49	25381
Relevant (can’t tell)	49	210
Relevant (opinionated)	13	159
Relevant (factual)	13	92
Relevant (official)	8	63
Relevant (personal)	8	118
Relevant (indepth)	18	220
Relevant (shallow)	18	176

Table 2: Breakdown of relevance levels for the faceted blog distillation task.

it only in passing (i.e. the blog is not principally about the target X).

1 *Relevant*. The blog has a clear principal, and recurring interest in the target X, but it is not relevant to either facet (or both facets).

2 The blog is relevant and is clearly inclined towards the “first” facet inclination (opinionated, personal, or indepth).

3 The blog is relevant and is clearly inclined towards the “second” facet inclination (factual, official, or shallow).

Participating groups were allowed to submit up to 4 runs for the faceted blog distillation task. TREC received a total of 29 faceted blog distillation runs from 9 groups, including 24 title-only runs, 3 title-description-narrative runs and 2 title-narrative runs. While all submitted runs were automatic, only 7 groups submitted title-only runs. NIST formed the pool by pooling all submitted runs (and all three rankings in each run) to depth 30. All assessments have been conducted by NIST assessors. Table 2 shows the breakdown of the relevance assessments of the pooled blogs per-facet, using the relevance levels described above. It is worth noting that 96% of the pooled blogs were judged as irrelevant. The assessors did not make use of the -1 relevance label, introduced to allow assessors to discard blogs if their associated blog posts URL were offensive. Indeed, all pooled blogs were judged.

As shown in Table 2, out of the 50 new topics, one topic did not have any associated relevant blogs in the pool (label 1). On the other hand, 10 topics did not have any relevant blog result for at least one facet inclination (label $>$ 1). Hence, in order that scores among the ‘none’ and faceted rankings are comparable, the reported official evaluation results only use 39 topics. These 39 topics have at least one relevant blog for each inclination of the facet (e.g. one relevant ‘indepth’ blog and one relevant ‘shallow’ blog).

3.3 Results

The blog distillation task is an adhoc-like search task. As a consequence, the primary measure for evaluating the retrieval performance of the participating groups is the mean average precision (MAP). Other metrics reported are R-Precision (rPrec), binary Preference (bPref), and Precision at 10 documents (P@10).

Table 3 provides the average best, and median MAP and P@10 measures for each topic and facet, across all submitted 29 faceted blog distillation runs. In general, the retrieval performances of the deployed participating systems have been average at best. This is somehow expected, given the statistics shown by the pool in Table 2, where the overwhelming majority of the retrieved blogs by the participating systems have been deemed irrelevant. While the obtained retrieval performances may reflect the intrinsic complexity and difficulty of the faceted blog distillation task, and the possible early-stage of the deployed faceted search approaches, it is unclear whether the level of difficulty of the actual new topics has further aggravated the obtained performances.

	Facet	MAP	P@10
Best	Baseline	0.3617	0.5308
Median		0.1285	0.2436
Best	Opinionated	0.2338	0.2615
Median		0.0727	0.1000
Best	Factual	0.2945	0.2308
Median		0.0685	0.0769
Best	Official	0.3167	0.2375
Median		0.0560	0.0625
Best	Personal	0.2995	0.3250
Median		0.0937	0.1125
Best	Indepth	0.3489	0.2778
Median		0.0549	0.0889
Best	Shallow	0.1906	0.2111
Median		0.0250	0.0333

Table 3: Best and Medians for the various facets of the faceted blog distillation task.

As mentioned in Section 3.2, the participating groups have been asked to submit for each topic, a ranking of blogs, where no facet inclination detection is applied, i.e. no particular faceted search approach is deployed, which is akin to a topic-relevance *baseline*. For the evaluation of the baseline rankings, all returned blogs judged 1 or above as per the assessment procedure described in Section 3.2 are deemed relevant. For the 39 retained topics, Table 4 shows the best-scoring baseline title-only automatic run for each group in terms of topic-relevance MAP, and sorted in decreasing order. The rPrec, bPref and P@10 measures are also reported. Two groups, namely IowaS and BIT, did not submit any title-only run. Table 5 shows the best automatic baseline run from each group, in terms of topic-relevance MAP, regardless of the topic length used.

Next, we show the results of the participating groups in faceted blog distillation search. Since different topics were assessed with respect to different facets, each run is evaluated by averaging its performance over all 39 topics, but with its performance on a particular topic calculated with respect to the first and second facet inclinations (relevance labels 2 and 3, respectively) appropriate to the topic. For example, for the topic 1103 (Opinionated), we assess the performance of the run on the ‘opinionated’ and ‘factual’ inclinations of the facet. More precisely, given that three facets were used in the topics, each run is assessed on its resulting associated 6 rankings (2 rankings per-facet, corresponding to each inclination of the facet). Table 6 selects the best automatic run for each group, which had the best overall *All MAP*. All MAP is calculated as the average of the AP for all of the queries for each opinion facet inclination. In other words, Table 6 shows the best deployed system per-group on average on all facets. Note again that two groups, IowaS and BIT, did not submit any title-only run.

Table 7 provides a summary of the results obtained by the four groups who achieved the best retrieval performances according to the MAP measure on a given facet inclination, i.e. MAP(facet). To assess the extent to which the faceted approach of a given run is effective, we compare its retrieval effectiveness on a given facet inclination (i.e. MAP(facet)) to the performance of the same run when no particular facet detection inclination approach is used (i.e. the effectiveness measure of the baseline ranking denoted by MAP-(baseline)). A relative MAP increase in performance indicates that the used faceted search strategy was successful. A relative MAP decrease in performance indicates that the deployed faceted search technique did not help in retrieval (see column *Improvement* in

Group	Run	MAP	P@10	bPref	rPrec
buptpris___2009	prisb	**0.2756**	**0.2767**	**0.3206**	**0.3821**
ICTNET	ICTNETBDRUN2	0.2399	0.2384	0.2863	0.3513
USI	combined	0.2326	0.2409	0.2815	0.3308
FEUP	FEUPirlab2	0.1752	0.1986	0.2447	0.3282
uogTr	uogTrFBAlr	0.1317	0.1531	0.2004	0.2333
UAms	IlpsBDm2T	0.0803	0.0966	0.1336	0.1590
knowcenter	nounfull	0.0624	0.0742	0.0980	0.1410

Table 4: Faceted blog-distillation task: Baseline ranking (i.e. no facet approach is applied), automatic title-only runs, 1 per group. Ranked by MAP, where relevant is blogs judged ≥ 1. The IowaS and BIT groups did not submit title-only runs.

Group	Run	Topic Fields	MAP	P@10	bPref	rPrec
buptpris___2009	pris	TDN	**0.2821**	**0.2852**	**0.3420**	**0.3949**
ICTNET	ICTNETBDRUN2	T	0.2399	0.2384	0.2863	0.3513
USI	combined	T	0.2326	0.2409	0.2815	0.3308
FEUP	FEUPirlab2	T	0.1752	0.1986	0.2447	0.3282
uogTr	uogTrFBAlr	T	0.1317	0.1531	0.2004	0.2333
BIT	BIT09PH	TDN	0.1165	0.1347	0.1714	0.2513
UAms	IlpsBDm2T	T	0.0803	0.0966	0.1336	0.1590
IowaS	IowaSBD0902	TN	0.0785	0.0978	0.1368	0.1564
knowcenter	nounfull	T	0.0624	0.0742	0.0980	0.1410

Table 5: Faceted blog-distillation task: Baseline ranking, 1 per group. Ranked by MAP, where relevant is blogs judged ≥ 1.

Table 7). It is worth noting that the MAP(baseline) for a given facet inclination (e.g. 'opinionated') is the evaluation of the baseline ranking when only the (e.g. 'opinionated') blogs are treated as relevant. This means that MAP(baseline) changes on a per-facet basis, and is not the same as the figures reported in Tables 4 and 5.

From the results in Table 7, we observe that in almost all cases, when the faceted search approaches are deployed, a decrease in performance is observed in comparison to the underlying baseline rankings. In fact, runs *FEUPirlab2-4* from FEUP (Universidade do Porto), which feature as top runs on various facet inclinations are all baseline-only runs that did not attempt any faceted search approach. Only 3 groups had runs which showed positive improvement on some facet inclinations: run *uogTrFBHlr* from the uogTr group (University of Glasgow) which deployed a faceted search task that improved the corresponding baseline on the 'factual' and 'official' facet inclinations; run *BIT09PH* by BIT (Beijing Institute of Technology) showed improvement by using facet-specific language models over a TDN baseline (i.e. a run that used all possible topic fields); and run *regularized* by USI (University of Lugano), which deployed a faceted retrieval strategy that improved the corresponding baseline ranking on the 'shallow' facet inclination. Overall, the obtained results show that the faceted blog distillation task has been particularly challenging to the participating groups.

3.4 Participants Approaches

In the following, we review the approaches of the participants. For more details, readers are referred to the proceedings papers of the various participants.

Most of the groups indexed only the permalinks component of the Blogs08 collection. The only exceptions are groups UAms (University of Amsterdam) and knowcenter (Know-Center), which only indexed the feeds component of the collection. It is of note that the UAms group only ran experiments on a title-only index of the Blogs08 feeds component. Finally, the group BIT (Beijing Institute of Technology) compared a permalinks-based index with another containing both the permalinks and homepages components.

For retrieval, many of the groups adopted a two-stage approach, where they first identified topic-relevant feeds, regardless of the facet inclination (baseline system). In the second stage, they use different classification or heuristic techniques to estimate the extent to which a retrieved blog is relevant to a facet inclination.

Almost all deployed retrieval techniques for the first stage (i.e. baseline ranking) scored a blog based on the scores of its corresponding relevant posts. In particular, uogTr (University of Glasgow) and UAms adapted their previously used expert search models to feed search. In addition, UAms used external query expansion on a news corpus and on Wikipedia to further enhance their baseline. The BIT group used a mixture of language models based on a global representation of the blogs, where a blog is treated as a virtual document composed of the concatenation of all its blog posts, again a document representation widely used in expert search. The FEUP group (Universidade do Porto) used a baseline run based on a BM25 ranking produced with the Terrier framework. For ranking the feeds, they focused on the temporal information available in most individual posts in Blogs08 collection to amplify (or reduce) each post's score before aggregating it into a feed score. Similarly, ICTNET (Institute of Computing Technology, Chinese Academy of Sciences) also ranked posts by BM25 before combining to rank blogs. The buptpris___2009 group (Beijing University of Posts and Telecommunications) used a basic topic relevance model, and for some runs, expanded the queries using terms from the description and narrative topic fields. The USI group (University of Lugano) experimented with two techniques for topic-relevance feed search. In the first approach, they used fuzzy aggregation methods for combining post relevance scores in each blog to calculate blog scores as a whole. In the second approach, they use regularisation methods for smoothing relevance scores based on the similarity between the retrieved blogs. They carry out regularisation on two types of scores: posts relevance scores and virtual document relevance scores (where each blog is represented by the concatenation of its most relevant posts). The IowaS group (University of Iowa) used a latent Dirichlet relevance model and query expansion using the Lucene framework. Finally, the knowcenter group ranked the top 100 topic-relevant blogs according to the accumulated relevance score of its relevant blog entries.

In the second stage, for the identification of the facet inclination of a given feed, the IowaS group used sentiment classifiers and

Group	Run	Topic Fields	MAP						
			All	Opinion	Factual	Official	Personal	Indepth	Shallow
USI	regularized	T	**0.1261**	0.0897	0.1044	0.1577	0.1337	0.1469	**0.1298**
FEUP	FEUPirlab2	T	0.1198	0.1068	0.1339	0.1523	0.1791	**0.1489**	0.0491
ICTNET	ICTNETBDRUN2	T	0.1030	**0.1259**	0.1176	0.0257	**0.1855**	0.1200	0.0567
BIT	BIT09PH	TDN	0.1026	0.0798	**0.1350**	0.1047	0.1239	0.1403	0.0475
uogTr	uogTrFBHlr	T	0.0918	0.0919	0.1103	**0.1965**	0.0739	0.1015	0.0301
buptpris_2009	prisb	T	0.0826	0.0719	0.0542	0.0672	0.0770	0.1362	0.0667
UAms	IlpsBDm2T	T	0.0534	0.0361	0.0391	0.0743	0.0795	0.0896	0.0194
knowcenter	punctfull	T	0.0459	0.0797	0.0382	0.0202	0.0996	0.0478	0.0125
IowaS	IowaSBD0902	TN	0.0453	0.0385	0.0804	0.0583	0.0174	0.0655	0.0111

Table 6: Faceted blog-distillation task: Best deployed faceted ranking systems on average on all facets, 1 per group. Ranked by All MAP. The IowaS and BIT groups did not submit title-only runs, and hence their best run (regardless of topic) is shown.

various heuristics for ranking posts according to each facet. The knowcenter group classified the topic-relevant blogs using a Support Vector Machine trained on a manually labelled subset of the TREC Blogs08 dataset. Three experiments were conducted, one based on nouns, one based on stylometric properties, and one based on punctuation statistics. They report that their facet identification approach was successful, although a significant number of candidate blogs were not retrieved at all (they only managed to successfully indexed 680k out of 1.3M blogs). The ICTNET group learned a classifier for the In-depth facet, while for other facets, a facet score was computed using facet term weights, which measured the extent to which the post is appropriate for a given facet inclination. The buptpris_2009 group used a Maximum Entropy-based classifier for the Opinionated facet, while the Personal facet was predicted based on the presence of named entities, and the In-depth facet was predicted based on post length. The UAms group used indicators such as post length for 'indepth', or first person pronouns for 'personal' to estimate the facet inclination of posts/blogs. It is of note that FEUP did not attempt any facet inclination identification, and submitted baseline-only rankings. In the following, we provide a detailed description of the three deployed faceted blog distillation methods that led to improvements over the baseline ranking system according to Table 7.

USI first generated positive and negative facet scores for each retrieved document and then combined the facet rankings with the relevance ranking using Borda Fuse. For the Indepth facet, they calculated the Cross Entropy (CE) between each retrieved document and the collection as a whole, using it as a the positive facet score since high CE indicates that the document contains many rare and informative words. Negated CE was used as the negative facet score. For the Opinionated facet, they built lexicons of opinionated and objective words from the Blogs06 collection using document frequency-based Mutual Information (MI) to weight terms. They calculated positive and negative facet scores for each retrieved document by averaging over the MI weights for each word in the document. Finally for the personal versus official facet, the same scores were used as for the Opinionated facet.

The uogTr group deployed machine learning techniques to identify blogs fulfilling the desired facet inclination from a baseline ranking produced by the Voting Model. In their first approach, different classifiers were trained to estimate the extent to which a given blog matched either inclination of a facet. In their second approach, the AdaRank learning-to-rank technique was used to learn a ranking model for each facet inclination. To enable their approaches, a large set of features – computed from both blog posts and entire blogs – and some training examples were produced.

The BIT blog retrieval system used a mixture of language models based on global representation. This model treats a blog as a big document where all postings of the blog are concatenated into a virtual document. In addition, the system uses a mixture of language models to construct the topic-facet language models. The topic-facet language model jointly models faceted words and topic words to rank blogs by both faceted relevance and topic relevance.

4. TOP NEWS STORIES TASK

A poll by Technorati found that 30% of bloggers considered that they were blogging about news-related topics [7]. Similarly, Mishne & de Rijke [8] showed a strong link between blog searches and recent news - indeed almost 20% of searches for blogs were news-related. As an illustration, Thelwall [12] explored how bloggers reacted to the London bombings, showing that bloggers respond quickly to news as it happens. Furthermore, both König et al. [4] and Sayyadi et al. [11] have exploited the blogosphere for event analysis and detection, showing that news events can be detected within the blogosphere.

On the other hand, on a daily basis, news editors of newspapers and news websites need to decide which stories are sufficiently important to place on their front page. Similarly, Web-based news aggregators (such as Google News) give users access to broad perspectives on the important news stories being reported, by grouping articles into coherent news events. However, deciding automatically on which top stories to show is an important problem without much research literature. Relatedly, in a given news article, some newspapers or news websites may provide links to related blog posts, often covering a diverse set of perspectives and opinions about the news story. These also may be hand selected, or automatically identified.

For these two scenarios, we have developed the top news identification task of the TREC 2009 Blog track. This task had two aims: firstly, to evaluate the ability of systems to automatically identify the top news stories on a given day, as an editor would do, but using only evidence from the blogosphere; secondly, to provide related blog posts covering diverse perspectives of that news story – to address the issue of which relevant blog posts a system needs to display to the users as an accompaniment to a given identified news story, so as to provide a good coverage of the different perspectives and aspects of the story. In the top news identification task, we use Blogs08 as a sample of the blogosphere. Blogs08 is particularly suitable, as it has a long timespan covering many important news events in 2008 (e.g. USA elections, China earthquake, etc).

4.1 Task Definition and Topics

To keep the difficulty of the task at a reasonable level during the TREC 2009 pilot study, we adopted a task that is more of a Retrospective Event Detection (RED) type [13], i.e. it uses the Blogs08

Group	Run	Topic Fields	MAP(baseline)	MAP(facet)	Improvement
Opinionated					
ICTNET	ICTNETBDRUN2	T	**0.1723**	**0.1259**	-26.93%
USI	OWA	T	0.1311	0.1176	-10.30%
FEUP	FEUPirlab3	T	0.1121	0.1121	0.00%
uogTr	uogTrFBMclas	T	0.1012	0.0988	-2.37%
Factual					
FEUP	FEUPirlab3	T	0.1370	**0.1370**	0.00%
BIT	BIT09PH	TDN	0.1331	0.1350	1.43%
ICTNET	ICTNETBDRUN2	T	**0.1389**	0.1176	-15.33%
uogTr	uogTrFBHlr	T	0.0954	0.1103	15.62%
Official					
USI	OWA	T	**0.2303**	**0.1973**	-14.33%
uogTr	uogTrFBHlr	T	0.1691	0.1965	16.20%
FEUP	FEUPirlab3	T	0.1589	0.1589	0.00%
BIT	BIT09PH	TDN	0.1064	0.1047	-1.60%
Personal					
USI	RegLDM	T	0.1548	**0.2169**	40.12%
ICTNET	ICTNETBDRUN2	T	**0.2049**	0.1855	-9.47%
FEUP	FEUPirlab2	T	0.1791	0.1791	0.00%
BIT	BIT09PH	TDN	0.1199	0.1239	3.34%
Indepth					
buptpris___2009	pris	TDN	**0.3124**	**0.1955**	-37.42%
FEUP	FEUPirlab4	T	0.1494	0.1494	0.00%
USI	regularized	T	0.1859	0.1469	-20.98%
BIT	BIT09PH	TDN	0.1392	0.1403	0.79%
Shallow					
USI	regularized	T	**0.1211**	**0.1298**	7.18%
buptpris___2009	prisb	T	0.1157	0.0667	-42.35%
ICTNET	ICTNETBDRUN2	T	0.0921	0.0567	-38.44%
FEUP	FEUPirlab1	T	0.0506	0.0506	0.00%

Table 7: For each facet, the best run from the top four groups by MAP(facet), sorted by MAP(facet). MAP(baseline) is the MAP of the baseline ranking for that facet inclination. FEUP did not attempt deploying any faceted search approach.

corpus as a static collection, where the participating systems can use any evidence from the whole Blogs08 collection. Next, the organisers obtained permission from the New York Times (NYT) to distribute a large sample of news headlines and their corresponding publication date. These headlines cover all articles published by NYT throughout the whole timespan of the Blogs08 corpus. Moreover, while the content of the articles is not included, the NYT URL corresponding to each headline was provided. It is of note that these URLs could be used to fetch the full-content of the articles from the NYT website.

In response to a given unit of time (the query date), the task requires the participating groups to provide a ranking of the top headlines that they think were important on the specified day. Moreover for each headline, they were asked to provide a ranking of supporting blog posts which are relevant to and discuss the news story headline. Finally, the blog posts selected for a given headline should be diverse in that they discuss different aspects, perspectives or opinions of the news story.

The dates of the provided headlines are the ones used by the news broadcaster (i.e. NYT in our case). For example, a story that happens in Europe very early in the morning of day d, can be issued with a date $d - 1$ by the American news broadcaster. Because of this possible time disparity between the date when the headline was issued by the news broadcaster and the one where the story actually happened, we have asked the participating systems to rank all headlines corresponding to the query date $d \pm 1$ days (i.e. headlines on day d, day $d - 1$, and day $d + 1$), so as to have a good grasp of the events that happened on day d. However, it is important to stress that this is not akin to judging all top headlines published on date d as being important for any date $d \pm 1$. Indeed, the reference date for an event (to assess relevance) is the date when the story actually happened (see Section 4.2).

Moreover, it is of note that relevant blog posts may naturally be posted on or after the date of the news headline, but even shortly before the provided headline date (recall the possible time disparity). Therefore, given the RED type of the pilot top stories identification task, these blog posts just have to be on topic, i.e. related to the news headline. In addition, the blog posts selected by the participating system for a given headline should be diverse in that they discuss different aspects, perspectives or opinions of the news story.

The organisers supplied 55 new topics, covering a wide range of global, political, financial, cultural, sports, and technology events that happened during the timespan of Blogs08, such as the Chinese Earthquake, President Obama's inaugural address, the banking/financial crisis, the Academy Awards, the Beijing Olympics, and the Microsoft-Yahoo aborted deal. Each topic corresponds to a date within the timespan of Blogs08, and does not provide additional description or narrative fields. An example of a topic illustrating the format of the topics is shown in Figure 2.

4.2 Assessments and Pools

Participating groups were allowed to submit up to four runs for the top stories identification task. Each run consists of a ranking of 100 headlines, and their corresponding supporting relevant posts.

```
<top>
<num>TS09-33</num>
<date>2008-08-25</date>
</top>
```

Figure 2: Blog track 2009, top stories identification, topic 33.

The required system responses are similar to the TREC Enterprise track Expert Search task format. It includes a list of supporting relevant discussive documents (at most 10) in the response covering various aspects of the news story. NIST received a total of 25 runs from 7 groups. All runs were automatic. The pool was formed by taking the top 20 headlines per topic from each submitted run, and the top 10 supporting documents for each pooled headline. As described in Section 4.1, only stories which were published ±1 day around the dates of interest were pooled.

The assessments were conducted by the participating groups using a newly developed judging interface. The assessment has two phases. In the first phase, the assessors were asked to judge the most important headlines for each query day. In essence, the assessors were asked to think like the *editor* of a newspaper or a news website. For each headline, they were asked to make a decision about whether the headline actually occurred on the query day, and whether they would have placed it on the front page of their news website or newspaper on that day. For each story, they should then select one of the following importance levels:

Not Important: This news story corresponding to the headline was not one of the most important that day.

Important: This story was one of the most important that day.

To take into account the time disparity (see Section 4.1), it was stressed to the assessors that they should only judge a headline as important if the event the headline is referring to actually happened on the day of the query. For instance, an aircraft disaster may occur on day X, but be reported by the NY Times on day $X + 1$ (due to reporting lag) or day $X - 1$ (because the story happened in a different part of the world). In this case, they should only judge the headline important for day X. The primary evaluation metric for the effectiveness of the top headlines identification is MAP.

We provided the assessors with several criteria to help them decide on the newsworthiness of an event (Timing, Significance, Prominence, Human Interest, Proximity)[2]. The assessors were free to judge a headline story based on the title and snippet provided in the judging user interface, or to follow the URL to the real NYT page. They were also permitted to use their recollection of events that happened on that day, or to use the Web or other resources when deciding what stories were important.

The second phase of the assessments examines how effective each system is at identifying relevant blog posts to each selected headline. In particular, the assessors were required to judge the relevant blog posts for the identified important headlines, and to group the relevant posts into various aspects of the news headline. This two-stage judgement procedure (first judge headlines, then judge blog posts) was devised to keep the relevance assessments workload reasonable. Indeed, the relevance assessments in phase 1 were very light and had the advantage of reducing considerably the relevance judgements workload in phase 2, as all irrelevant headlines and associated blog posts were discarded. To this end, and for a fair comparison of the best performing systems, the initially formed pool of blog posts was trimmed to only those posts that are associated to relevant headlines which were retrieved by at least 7 of the best 9 performing headline ranking runs, as ranked by MAP. For each blog post in this new pool, the assessor was required to read the post, and decide if it is relevant to the headline. There were two relevance options:

Relevance Level	# Stories
Not Important	9453
Important	1434

Table 8: Breakdown of relevance levels for the top news story identification task, headline judgements.

Relevance Level	# Blog Posts
Not Relevant	3453
Relevant	4375

Table 9: Breakdown of relevance levels for the top news story identification task, blog post judgements.

Not Relevant: This blog post has no bearing on the news story.

Relevant: This blog post discusses an aspect of the news story.

If the blog post was deemed relevant, then the new judging interface provides support for the assessor to select an existing aspect that describes the aspect of the news story that the post covers/discusses/addresses, or to enter a new aspect. For example, say the headline concerns the Obama victory announcement on 5th November. By judging blog posts, the assessors may identify aspects such as “Factual reporting”, “Analysis of win” and “Transition period opinions”. The assessment of the extent to which the supplied blog posts by a participating system are diverse are measured using the α-nDCG [2] or IA-Precision [1] metrics, in a fashion similar to the Web track 2009 diversity task. Note however, that unlike the Web track, the subtopics/perspectives are not predefined, as they are identified by the assessors after pooling, during the phase 2 assessment stage.

4.3 Results

First, we provide an overview of the results of the first stage of the top stories identification task, namely the effectiveness of the participating systems in identifying the top headlines for a given query date. The NY Times headlines corpus includes 101,730 news headlines in total, with 242 headlines as the mean number of stories per day. Table 8 provides a distribution of relevance levels in the formed pool of headlines for all 55 query dates. In particular, about 86% of the headlines in the pool were not deemed to be important by the assessors. This was somewhat reduced, as we asked some assessors to reduce the number of important stories they had assessed for each day.

We sampled 258 judged important headlines for which to perform blog post judging – in particular, important headlines retrieved by at least 7 of the top 9 runs by MAP were assessed. 258 headlines represented a tradeoff between collection reusability and judging effort. Indeed, these 258 headlines resulted in a pool of 8225 blog posts that were to be assessed – an average of 32 blog posts per relevant headline. Table 9 shows the number of relevant and not relevant blog posts for the 258 headlines. From the results, we note that identifying relevant posts was fairly straightforward, with 56% of pooled blog posts being relevant to the headlines. Moreover, to handle the diversity element of the task, assessors grouped relevant headlines into different aspects. On average, 4.5 aspects were identified for each headline, suggesting that assessors were able to form relevant blog posts into a few coherent aspect groupings.

[2] See: `http://www.mediacollege.com/journalism/news/newsworthy.html` for more details.

	MAP	P@10
Best	0.2553	0.4873
Median	0.0445	0.1164

	α-NDCG@10	IA-P@10
Best	0.7723	0.2759
Median	0.0217	0.0041

Table 10: Best and medians for the headline ranking and diverse blog post ranking parts of top news stories identification task.

In all, 25 runs from seven groups were submitted to the top news stories identification task. While groups were permitted to use external evidence in their runs, these were to be ranked separately. However, in the submitted runs, no groups made use of external evidence.

For the headline ranking element of the task, the top-half of Table 10 provides the average best and median MAP and P@10 effectiveness measures for each topic, across all submitted 25 runs to the top news stories identification task. The reported figures are particularly low, suggesting that most submitted runs had difficulties in producing an effective ranking of headlines.

Table 11 shows the best scoring top headlines ranking run for each group, ranked by decreasing MAP. The P@5, P@10, MRR and bPref effectiveness measures are also reported. In general, the performances of the submitted runs show that there is still a large room for improvement towards achieving effective top headlines identification and ranking techniques.

Next, we assess the ability of the runs to retrieve relevant diverse blog posts. The bottom-half of Table 10 provides the best and median α-NDCG@10 ($\alpha = 0.5$), and IA-P@10 measures of the participants' runs, for each of the 258 headlines that had blog posts assessed. The marked difference between the best and median suggests that many systems struggled in obtaining good performance with this part of the task, probably due to poor headline ranking performance. Table 12 shows the best scoring diverse blog posting runs for each group. The table is ranked by (mean) α-NDCG@10 ($\alpha = 0.5$), while the mean of α-NDCG@5, IA-P@5 and IA-P@10 measures are also reported. Note that different runs retrieved different numbers of important headlines. Therefore, the reported mean values are calculated over all 258 assessed headlines, and hence are somewhat correlated with the runs' performance on the top headline identification element of the task.

Finally, we examined if measuring the ability of the systems to retrieve diverse blog posts (as measured by α-NDCG@10) was noticeably different from their ability to just retrieve relevant blog posts (as measured by MAP). Across all 25 submitted runs, there was a correlation of Spearman's $\rho = 0.984$, Kendall's $\tau = 0.906$, suggesting that the rankings were very similar overall.

4.4 Participants Approaches

In the following, we review the approaches of the participants in the top news stories task. For more details, readers are again referred to the proceedings papers of the various participants.

In terms of indexing, five participating groups used a system that only indexed the permalinks component of the Blogs08 collection. In contrast, the UAms group (University of Amsterdam), which submitted 4 runs using an index of documents (blog posts) extracted from English-only blog feeds (i.e. the feeds component of Blogs08). Moreover, the USI group (University of Lugano), submitted a single run using all three components of the Blogs08 collection, namely feeds, permalinks and homepages.

In terms of retrieval models, uogTr (University of Glasgow) and UAms (University of Amsterdam) used techniques inspired by their expert search model to rank the headlines. The IowaS group (University of Iowa) ranked headlines using URL frequency-based ranking and similarity based on the latent Dirichlet relevance model, as well as query expansion. They did not diversify the ranking of blog posts. The shakwat group (University of Paris 8) experimented with a random-walk approach using a space built using semantic indexing, and containing the blog posts, as well as the headlines, in a window around the date of the topic. ICTNET (Institute of Computing Technology, Chinese Academy of Sciences) accumulated the BM25 scores for a given headline from the blog posts published that day, and were inspired by topic-focused text summarisation to build diverse blog post rankings. Finally, the USI group (University of Lugano) used an approach that ranks clusters of blog posts with respect to size and timespan. Below, we provide detailed descriptions of the methods and retrieval approaches deployed by the top performing groups.

The uogTr group explored an approach based on the Voting Model for expert search, hypothesising that the number of blog posts mentioning a headline (aka votes) is a good indicator of the importance of each headline. This allowed the most important headlines each day to be identified and scored. Investigating the bursty nature of the blogosphere, they further refined the headline scores through boosting those headlines which continued to be discussed in the blogosphere after the query date. The latter approach obtained a slight improvement over the baseline performance. Blog posts for the top headlines were retrieved using a hypergeometric weighting model from the Divergence from Randomness framework (DPH), while blog post diversification was achieved through the use of temporal distance or Maximal Marginal Relevance between blog posts.

The POSTECH_KLE group (KLE, Pohang University of Science & Technology) estimated the importance of a news headline for a date by linearly combining two probabilities. One is the probability that each news headline generates a given query date, calculated using feed-based or cluster-based approaches. The second is the prior probability that a news headline will be a top story for a given date, estimated using either time-based or term-based evidence.

The UAms group explored two approaches for identifying top stories: (i) news to blogs, and (ii) blogs to news, and both approaches are applied to a post index and a title-only index. The first approach uses an expert finding model and tries to calculate a headline likelihood: the probability of a headline given a date, where the date model is constructed using blog posts for that date. The second approach is more general, and tries to identify emerging topics from the blog posts of a given date. The most distinguishing terms are selected and clustered to form topics. In the final step, these term clusters are used as query on a headline index.

Of note is that two groups (uogTr and UAms) used approaches based on their existing expert search models. In the case of uogTr, this proved to be very effective.

5. CONCLUSIONS

The Blog track in its forth year has been markedly revamped, with the introduction of refined and typical search task scenarios that go beyond simple topical relevance or adhoc retrieval. In addition, the Blog track 2009 has seen the creation of a new and up-to-date sample of the blogosphere, Blogs08, which is one order of magnitude bigger than the older Blogs06 collection.

In TREC 2009, most of the organisers' efforts have been spent on defining the new search tasks, and on building an appropriate methodology and infrastructure to evaluate the effectiveness of the submitted runs. On the other hand, the participating groups have put significant efforts towards deploying appropriate indexing and retrieval strategies in line with the difficulties of the new tasks introduced in this year's revamped Blog track. The results on both

Group	Run	MAP	P@5	P@10	MRR	bPref
uogTr	uogTrTStimes	**0.1862**	**0.3236**	**0.3127**	**0.5390**	**0.2113**
POSTECH_KLE	KLEClusPrior	0.1605	0.2836	0.2964	0.4553	0.1930
UAms	IlpsTSExP	0.1354	0.2655	0.2745	0.4271	0.1813
IowaS	IowaSBT0904	0.0882	0.1600	0.1709	0.3294	0.1245
ICTNET	ICTNETTSRun1	0.0391	0.0800	0.0982	0.1801	0.0656
shakwat	ri1025rw5432	0.0388	0.1018	0.1200	0.2127	0.0725
USI	runtag	0.0062	0.0364	0.0182	0.1818	0.0062

Table 11: Top stories identification task: Ranking of runs for identifying important headlines, one run per group. Ranked by MAP.

Group	Run	α-NDCG@5	α-NDCG@10	IA-P@5	IA-P@10
uogTr	uogTrTSbmmr	**0.499**	**0.518**	**0.185**	**0.168**
POSTECH_KLE	KLEFeedPrior	0.490	0.504	0.178	0.162
IowaS	IowaSBT0901	0.328	0.341	0.117	0.099
UAms	IlpsTSExT	0.100	0.104	0.029	0.030
ICTNET	ICTNETTSRun1	0.066	0.073	0.027	0.024
shakwat	ri1025rw5432	0.003	0.002	0.001	0.000
USI	runtag	0.000	0.000	0.000	0.000

Table 12: Top stories identification task: Ranking of runs for identifying diverse blog posts, one run per group. Ranked by α-NDCG@10.

tasks confirm the complexities of the newly introduced task, and show that there is still a large scope for further research and improvement towards achieving effective retrieval strategies for both faceted blog search and top stories identification.

For TREC 2010, using lessons learnt from the current edition of the track, we will continue investigating the faceted blog distillation and the top stories identification tasks, with the introduction of various refinements, intended to facilitate research into considering the blogosphere as a time stream, instead of a static collection. For example, to keep the difficulty of the top story identification task at a reasonable level during the TREC 2009 pilot study, we adopted a task that is more of a Retrospective Event Detection (RED) type, i.e. it uses the Blogs08 corpus as a static collection. Instead, for TREC 2010, we will re-run the task considering the Blogs08 corpus as a time stream, i.e. as a New Event Detection task (NED). This is a more practical setting, as the headlines and the blog posts are ranked at a given time t, without information from/about the future.

Acknowledgements

The description of system runs are based on paragraphs contributed by the participating groups. We would like to express our thanks and appreciation to the New York Times (`http://www.nytimes.com/`), for providing the sample of headlines used in the top stories identification task. They are provided to support research in the TREC blog track. Thanks are also due to the 7 groups who assessed this year's top stories identification task topics.

6. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, S. Leong. Diversifying Search Results. In *Proceedings of WSDM-2009,* Barcelona, Spain, 2008.

[2] C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, Stefan Büttcher, and I. MacKinnon Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of SIGIR-2008,* Singapore, Singapore, 2008.

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

[4] A. C. König, M. Gamon, and Q. Wu. Click-through prediction for news queries. In *Proceedings of SIGIR-2009,* Boston MA, USA, 2009.

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

[6] C. Macdonald and I. Ounis. The TREC Blogs06 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`

[7] J. McLean. State of the Blogosphere, introduction, 2009. `http://technorati.com/blogging/article/state-of-the-blogosphere-2009-introduction`.

[8] G. Mishne, M. de Rijke. A Study of Blog Search. In *Proceedings of ECIR-2006,* London, UK, 2006.

[9] 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.

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

[11] H. Sayyadi, M. Hurst, and A. Maykov. Event detection and tracking in social streams. In *Proceedings of ICWSM 2009,* San Jose CA, USA, 2009.

[12] M. Thelwall. Bloggers during the London attacks: Top information sources and topics. In *Proceedings of the 3rd International Workshop on the Weblogging Ecosystem (WWE 2006)*, 2006.

[13] Y. Yang, T. Pierce, J.G. Carbonell A Study on Retrospective and On-line Event Detection. In *Proceedings of SIGIR-1998,* Melbourne, Australia, 1998.