Overview of the TREC-2010 Blog Track

Iadh Ounis, Craig Macdonald
University of Glasgow
Glasgow, UK
{iadh.ounis,craig.macdonald}@glasgow.ac.uk

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

1. INTRODUCTION

The Blog track aims to investigate the information seeking behaviour in the blogosphere. The track was initiated in 2006, and has used an incremental approach in tackling several search tasks by their level of difficulty. In TREC 2010, the track has investigated two main search tasks:

- *Faceted blog distillation*: A blog search task where systems aim to retrieve bloggers (i.e. RSS feeds) that have a recurring and central interest in a topic X [6], and which also satisfy a number of facets (or attributes), representing the nature or the quality of the sought blogs (e.g. opinionated, factual) [7].
- *Top stories identification*: A task that addresses news-related issues on the blogosphere, namely investigating whether the blogosphere can be leveraged to identify the top news stories of a given day in a *real-time* fashion. The task has also a search diversity flavour, where for a given story, a representative set of blog posts discussing the story from various perspectives [7] is shown to the user.

Both tasks this year used the Blogs08 corpus [7, 9], which is a sample of the blogosphere covering a timespan ranging from January 2008 to February 2009. The Blogs08 collection consists of roughly 1.4M blog feeds and 29M blog posts. In addition, for the purposes of the top stories identification task, a new large corpus of news stories covering the same timespan as Blogs08 has been released by Thomson Reuters. The corpus, called Thomson Reuters Research Collection (TRC2), contains both the headlines and content of over 1.8M news stories.

The topics for the faceted blog distillation task have been developed and assessed by NIST. On the other hand, for the top stories identification task, a number of dates have been sampled from the range of dates covered by Blogs08 and used as *query dates*. To develop topics for the search diversification component of the top stories identification task, the organisers have selected a set of news stories, for which the participating groups were asked to rank diverse blog posts discussing these stories in the blogosphere. In a marked departure from the usually adopted community judgements, in TREC 2010, the Blog track organisers made a first attempt at using crowdsourcing within TREC, where all runs submitted to the top stories task have been assessed through the use of the Amazon Mechanical Turk (AMT) service.

A total of 16 different groups participated in the 2010 Blog track, spread across four continents. Many groups attempted both tasks, deploying varying approaches ranging from advanced probabilistic retrieval models, to classification and/or machine learning-driven techniques. The remainder of this paper is structured as follows. Section 2 describes the faceted blog distillation task, and discusses the main obtained results by the participating groups. Section 3 describes the top stories identification task and its corresponding results. Concluding remarks are provided in Section 4.

2. FACETED BLOG DISTILLATION TASK

The blog distillation task was first introduced in TREC 2007 [6]. 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 [6, 9, 14]. However, in its TREC 2007 & 2008 incarnations, the blog distillation task only focused on topical relevance. It did not address the *quality* aspect or the nature of the retrieved blogs.

Inspired by a position paper by Hearst et al. [4] in TREC 2009 [7], we proposed a refinement of the blog distillation task that takes into account a number of facets that allow the filtering of blogs according to various attributes, such as the authority of the blog, its opinionated nature, the trustworthiness of its authors, or the genre of the blog and its style of writing.

As detailed in [7], the faceted blog distillation task mimics an exploratory search task. 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. 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. 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 and nature of the blogs is characterised through a set of facet *inclinations*.

2.1 Task Definition and Topics

The same three facets proposed for the TREC 2009 blog distillation task [7] have been used again in TREC 2010, all assumed to have binary inclinations for operational simplicity. In particular, the three facets used for TREC 2010 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

```
<top>
<num> Number: 1154 </num>

<query> chinese economy </query>

<desc> Description:
I am interested in blogs on the
Chinese economy.
</desc>

<facet> opinionated </facet>

<narr> Narrative:
I am looking for blogs that discuss
the Chinese economy.  Major economic
developments in China are relevant,
but minor events such as factory
openings are not relevant.  Information
about world events, or events in other
countries is relevant as long as the
focus is on the impact on the Chinese
economy.
</narr>

</top>
```

Figure 1: Blog track 2010, faceted blog distillation task, topic 1154. The query tag corresponds to the traditional topic title.

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 analyses 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).

NIST has developed 50 new topics for TREC 2010. During the topic development, one appropriate facet was chosen for each topic. In particular, the Opinionated facet has been associated to 17 topics, the Personal facet has been associated to 16 topics, and the In-depth facet has been associated to 17 topics. An example of a topic associated with the Opinionated facet is included in Figure 1.

A fundamental objective for the TREC Blog track 2010 faceted blog distillation task was to identify the most effective and robust ranking approaches with respect to a given facet. As a consequence, inspired by the experimental setup used for the TREC 2008 opinion-finding task [14], in 2010, the faceted blog distillation task involved two separate sub-tasks:

- *Baseline Blog Distillation:* This sub-task consists in ranking 100 blogs that the deployed system assesses to be relevant to a topic, without any consideration of the facet attached to this topic. This task exactly corresponds to the TREC 2007 & 2008 blog distillation tasks [6, 14], or the "None" facet rankings from TREC 2009 [7].

- *Faceted Blog Distillation:* In this sub-task, for each topic, systems should supply two rankings of 100 blogs each: one for the first inclination of the facet enabled, and one with the second inclination of the facet enabled. For example, for the Personal facet, the first ranking would have 100 blogs that the system assesses to be 'personal', and the second ranking would have 100 blogs, which the system assesses to be 'official'.

To aid cross-comparison and to allow participants to study the performance of their specific faceted search approach across a range of different baseline systems, NIST selected three *"standard baselines"* from the submitted baseline blog distillation runs, which were redistributed to all participants prior to the faceted blog distillation sub-task submission deadline.

Finally, to permit the future analysis of the difficulty of the topics across the years, as well as to facilitate the investigation of the effect of various training regimes, the participating groups were asked to submit their runs using the 50 new TREC 2010 topics and the 50 old topics from the TREC 2009 faceted blog distillation task.

2.2 Assessments and Pools

Participating groups were allowed to submit up to 2 runs to the baseline blog distillation sub-task, including a compulsory automatic *query-only* run. They were then permitted to submit up to 4 runs, which are based on each of their two previously submitted baseline runs (i.e. 4 runs per own baseline, 8 maximum). One of these submitted runs must be an automatic, query-only run. Moreover, to improve the quality of the pool, we encouraged groups to submit manual runs, and to avoid varying the length of the query (with/without description or narrative) from the baseline to the faceted runs, so as to ensure the clarity of their analysis.

In addition, groups could submit up to 4 runs for each of the provided 3 standard baselines. Hence, in total, each group could submit a maximum of 20 runs (4* (3 standard baselines + 2 own baseline runs)). To aid the cross-comparison of the deployed faceted ranking approaches and to facilitate the analysis of their performance and robustness, the participating groups were encouraged to apply any given facet ranking approach on each of the three standard baselines. For those runs where the system cannot be clearly broken down into baseline and facet-ranking features, the groups were advised to indicate "N/A" as the baseline run.

Based on observation from the TREC 2009 faceted blog distillation task, where there was no noticeable reduction in the quality of the test collection when only pooling from the baseline runs, the TREC 2010 pool was drawn only from the submitted baseline runs rather than from all baselines and faceted search runs. All baseline runs were pooled to depth 40. Similar to TREC 2009 [7], 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 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').

Relevance Level	# Queries	# Blogs
Not Relevant	31	7276
Relevant (can't tell)	31	88
Relevant (opinionated)	7	208
Relevant (factual)	7	68
Relevant (official)	10	86
Relevant (personal)	10	119
Relevant (indepth)	14	103
Relevant (shallow)	14	181

Table 1: Breakdown of relevance levels for the faceted blog distillation sub-task.

	MAP	P@10
Best	0.4340	0.6000
Median	0.1925	0.3097

Table 2: Best and Medians for the baseline blog distillation sub-task.

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

All assessments have been conducted by NIST assessors. The current evaluation results are preliminary as some topics do not yet have complete judgments. Of the 50 new topics, 31 have at least one relevant blog for each inclination of the topics' facet. Thus, for the purposes of analyses, all results reported below are for those 31 topics only.

For the 31 used topics, Table 1 shows the breakdown of the relevance assessments of the pooled blogs per-facet, using the relevance levels described above. About 90% of the pooled blogs were judged as irrelevant (a slight difference from the 96% irrelevant blogs found in the TREC 2009's pool).

In the following, Section 2.3 summarises the main obtained results by the participating groups on the 31 used new topics in the baseline blog distillation sub-task, while Section 2.4 provides an overview of the main results and findings from the corresponding faceted blog distillation sub-task.

2.3 Baseline Blog Distillation Results

As mentioned in Section 2.1, the baseline blog distillation sub-task is an adhoc task, where no particular faceted search approach is applied. This is akin to a *topic-relevance baseline*, where all returned blogs judged 1 or above as per the assessment procedure described in Section 2.2 are deemed relevant. 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 2 reports the per-topic best and medians of the submitted baseline blog distillation runs.

A total of 24 runs were submitted by 13 groups to the baseline blog distillation sub-task, of which there was 1 manual run. Table 3 shows the best submitted automatic query-only baseline blog distillation run from each participating group, ranked by MAP. Table 4 shows the best performing baseline run from each participating group, regardless of topic type and run type.

The top performing baseline run was submitted by the BIT group. They treated a blog as a large document where all postings of the blog are concatenated into a virtual document. They then used a language modelling approach to rank the resulting virtual documents. The PKUTM and HIT_LTRC groups also deployed a language modelling approach to aggregate blog post scores into blog scores. The ICTNET group used an approach based on ensemble ranking, while the uogTr group used an advanced version of their voting model-based approach. The PRIS group also used an approach based on the voting model. Finally, the StanfordNLP system used a probabilistic model that leverages individual blog post evidence to improve blog search.

From the 24 submitted baseline runs, NIST selected three standard baselines of varying performances, and made them available to all participating groups. Table 5 lists the three selected standard baseline runs, as well as their performances.

	Facet	MAP	P@10
Best	opinionated	0.4805	0.5429
Median		0.1275	0.2286
Best	factual	0.5452	0.3429
Median		0.1259	0.1143
Best	official	0.5181	0.4100
Median		0.1561	0.1300
Best	personal	0.4024	0.3900
Median		0.0827	0.1200
Best	indepth	0.5043	0.3071
Median		0.1408	0.1143
Best	shallow	0.2941	0.3571
Median		0.0712	0.1000

Table 6: Best and Medians for the various facets of the submitted faceted blog distillation runs.

2.4 Faceted Blog Distillation Results

In this section, we summarise the results of the participating groups in the faceted blog distillation sub-task. Since different topics were assessed with respect to different facets, each run is evaluated by averaging its performance over all used 31 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 1154 (Opinionated), we assess the performance of the run on the 'opinionated' and 'factual' inclinations of the facet. More precisely, similar to TREC 2009 [7], 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).

A total of 119 runs were submitted by 11 groups to the faceted blog distillation task. Of these, 70 were based on one of the three standard baselines, and 2 runs were manual. Table 6 reports the per-topic best and median results for each facet inclination, across all submitted faceted blog distillation runs. Similar to TREC 2009, the median performances varied from a facet to another, with the In-depth facet ('indepth' and 'shallow' inclinations) seemingly the most difficult.

Table 7 selects the best run for each group, which has the best overall *Mean Facet MAP*, regardless of topic type, used baseline (own or standard), or run type (automatic or manual). Mean Facet MAP is calculated as the mean of Facet MAP over all facet inclinations. In other words, Table 7 shows the best deployed system per-group on average on all facet inclinations.

Table 8 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. Facet MAP (facet run), regardless of topic length, baseline, or run type. 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. Facet MAP (facet run)) to the facet performance of the corresponding baseline run, which applies no particular facet inclination approach (denoted Facet MAP(baseline run)). For instance,

Group	Run	MAP	P@10	bPref	rPrec
BIT	BITblog10bl1	0.3501	0.3409	0.3970	0.4774
ICTNET	ICTNETBDRun2	0.3197	0.3394	0.3892	0.4484
PKUTM	PKUTMB1	0.2543	0.2569	0.3167	0.3581
HIT_LTRC	hitQuerybl	0.2495	0.2468	0.2917	0.3258
PRIS	pris	0.2208	0.2287	0.2878	0.3452
ULugano	bloggerModel	0.2054	0.2183	0.2670	0.3645
uogTr	uogTrapeMN5k	0.2024	0.2009	0.2519	0.3194
UICIR	uicfeedir1	0.1961	0.1888	0.2435	0.3097
PCUHK	PULM	0.1879	0.1903	0.2430	0.2839
feup	FEUPirlab2	0.1655	0.1827	0.2465	0.3161
RMIT	rmitprob	0.1330	0.1619	0.2128	0.2387
StanfordNLP	stanford2	0.1243	0.1428	0.1892	0.2129
UniNE	Run1	0.0394	0.0476	0.0685	0.0903

Table 3: Baseline blog distillation sub-task: automatic query-only runs, 1 per group. Ranked by MAP, where relevant is blogs judged ≥ 1.

Group	Run	Topic Fields	MAP	P@10	bPref	rPrec
BIT	BITblog10bl1	Q	0.3501	0.3409	0.3970	0.4774
ICTNET	ICTNETBDRun2	Q	0.3197	0.3394	0.3892	0.4484
HIT_LTRC	hitTDNbl$*$	QDN	0.2692	0.2611	0.3090	0.3613
PKUTM	PKUTMB1	Q	0.2543	0.2569	0.3167	0.3581
PRIS	pris	Q	0.2208	0.2287	0.2878	0.3452
ULugano	bloggerModel	Q	0.2054	0.2183	0.2670	0.3645
uogTr	uogTrapeMN5k	Q	0.2024	0.2009	0.2519	0.3194
UICIR	uicfeedir1	Q	0.1961	0.1888	0.2435	0.3097
PCUHK	PULM	Q	0.1879	0.1903	0.2430	0.2839
feup	FEUPirlab2	Q	0.1655	0.1827	0.2465	0.3161
RMIT	rmitprob	Q	0.1330	0.1619	0.2128	0.2387
StanfordNLP	stanford2	Q	0.1243	0.1428	0.1892	0.2129
UniNE	Run1	Q	0.0394	0.0476	0.0685	0.0903

Table 4: Baseline blog distillation sub-task: 1 per group. Ranked by MAP, where relevant is blogs judged ≥ 1. $*$ denotes a manual run.

Facet MAP(baseline run) 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 Facet MAP(baseline run) is different for each inclination, and is not the same as the figures reported in Tables 3 and 4. Increases are only reported when the facet runs did not report "N/A" as their corresponding baseline run. 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 Table 8). In general, the results show that the best performing runs for each inclination were able to improve over their corresponding baseline. In particular, promising improvements were achieved by the best performing runs for the 'opinionated', 'personal' and 'indepth' inclinations. For the 'factual', 'official' and 'shallow' inclinations, smaller margins of improvements were observed in the strongest runs.

Furthermore, we investigated the performance and robustness of a given faceted search approach across all the three provided standard baselines. The more a faceted search approach consistently improves the corresponding faceted retrieval performance of the three provided baselines, the more likely that it is effective and robust. For a fair comparison of the deployed faceted search approaches, we only considered the groups who attempted their faceted search approaches on all and each three provided standard baselines. Table 9 lists the best faceted search approach from each group. The Mean Facet MAP over all facet inclinations is reported for each standard baseline. Approaches are ranked by the average Mean Facet MAP over all three standard baselines. Increases in Mean Facet MAP per-standard baseline are also shown.

The results in Table 9 show that only one approach, namely the 'hitFeeds_' approach from the HIT_LTRC group, has consistently improved upon the faceted performances of the three provided standard baselines. In particular, the HIT_LTRC group used a Maximum Entropy Model toolkit to predict the facet inclination of every blog post in a feed. The ULugano group continued deploying their last year's approach based on scoring facets using cross entropy and various tailored lexicons, while the BIT group used SVM facet classifiers as input to a mixture of topic relevance model and facet relevance model constructed by pseudo-relevance feedback, respectively. The uogTr group used a learned voting approach combining over 900 post-level and blog-level features, including lexicons for each facet inclination. The UICIR group used concept-based retrieval to improve recall, and SVM classifiers to detect facets from these concepts.

3. TOP NEWS STORIES TASK

The top stories identification task was first run as a pilot task in TREC 2009 to address the news dimension of the blogosphere as detailed and motivated in [7]. In particular, it addresses whether the blogosphere can be used to identify the most important news stories for a given day. The task involves two aspects:

- Identifying top news stories for a given unit of time and category - the *Story Ranking Task*.

Std. Baseline	Baseline Run	Baseline MAP	Mean Facet MAP	MAP by Facet					
				opinionated	factual	official	personal	indepth	shallow
stdbaseline1	ICTNETBDRun2	0.3197	0.2082	0.2598	0.2693	0.2439	0.1377	0.2345	0.1038
stdbaseline2	uogTrapeMN5k	0.2024	0.1397	0.1054	0.2068	0.1938	0.0755	0.1309	0.1259
stdbaseline3	FEUPirlab1	0.1597	0.1170	0.0767	0.1660	0.2014	0.0899	0.0756	0.0923

Table 5: Performances of the standard baseline runs

Group	Run	Baseline	Topic Fields	Mean Facet MAP	MAP by Facet					
					opinionated	factual	official	personal	indepth	shallow
BIT	BIT10bl1fd3	BITblog10bl1	Q	0.2537	0.2415	0.2948	0.3301	0.1736	0.3211	0.1610
ICTNET	ICTNETFBD3	N/A	Q	0.2285	0.2554	0.2670	0.3134	0.1321	0.3042	0.0988
ULugano	LexMIRuns1	stdbaseline1	Q	0.2180	0.2656	0.2693	0.2415	0.2121	0.2365	0.0832
HIT_LTRC	hitFeeds1	stdbaseline1	?	0.2089	0.2607	0.2695	0.2464	0.1385	0.2349	0.1035
PKUTM	PKUTM121onB1	PKUTMB1	Q	0.1857	0.2807	0.1399	0.1930	0.1636	0.2398	0.0973
uogTr	uogTrfL919s1	stdbaseline1	Q	0.1837	0.2440	0.1369	0.2456	0.1017	0.2578	0.1162
UICIR	uicfbdstd1b	stdbaseline1	Q	0.1588	0.1938	0.1494	0.1948	0.1215	0.1917	0.1017
UniNE	run3swnpn10	stdbaseline1	QN	0.1434	0.1590	0.0929	0.2414	0.1293	0.1627	0.0752
PRIS	PrisStdQE1	stdbaseline1	QDN	0.1253	0.2065	0.2464	0.0052	0.0188	0.1990	0.0757
PCUHK	Std1stPI	stdbaseline1	Q	0.1006	0.1504	0.0930	0.1535	0.0789	0.0916	0.0362
RMIT	rmitfaceted	rmitprob	Q	0.0530	0.0682	0.0246	0.0515	0.0668	0.0662	0.0405

Table 7: Faceted blog-distillation sub-task: Best deployed faceted ranking systems on average on all facets, 1 per group. Ranked by Mean Facet MAP. Run hitFeeds1 did not declare its used topic fields.

- Identifying relevant blog posts for a given news story, that cover different/diverse aspects or opinions - the *News Blog Post Ranking Task*.

Differently from TREC 2009, the top stories identification task involved using a set of five standardised news categories (World, US, Sport, Science & Technology, Business) and, more importantly, was defined as an online event detection [18], i.e. it mimics a *real-time search* environment. To allow the components of participating systems to be evaluated independently, the task involved two stages: in the first stage, the participating groups aim to identify the top news stories for a given day. Once this task is completed, in the second stage, using a common set of top stories, the participating systems aim to identify and rank a *diverse* set of blog posts discussing each story.

3.1 Task Definition and Topics

In addition to the Blogs08 corpus, the participating groups were provided with a large new sample of news stories from throughout the timespan of the Blogs08 corpus. For the TREC Blog track 2010, Thomson Reuters has released the TRC2 newswire corpus, which contains both the headlines and content of over 1.8M news stories, and is distributed by NIST free of charge. The TRC2 corpus replaced the smaller New York Times (NYT) headline corpus used in TREC 2009.

As mentioned above, a further change from TREC 2009 is the use of categories, where the participating systems were asked to identify the top stories for a given category. In TREC 2010, the following five categories were used from a United States' perspective:

- *World* - all international news, including political news outside of USA.
- *U.S.* - all general United States news, including politics.
- *Sport* - all sport news.
- *Sci.Tech* - all technology/IT news as well as science/environment etc.

```
<DOC>
<DOCNO>TRC2-date-number</DOCNO>
<BLOGS08DAY>5</BLOGS08DAY>
<DATE>date</DATE>
<HEADLINE>headline of article</HEADLINE>
<CONTENT>content of article</CONTENT>
</DOC>
```

Figure 2: Blog track 2010, story ranking task. Format of a news article in the TRC2 collection.

- *Business* - all finance/economics/business news.

Importantly, as stressed above and differently from TREC 2009, the top story identification task was treated as an online event detection, thereby enforcing a real-time search scenario. To facilitate this, the organisers provided common timestamp information for each story (i.e. headline + content) in the TRC2 corpus, each blog post in the Blogs08 corpus, and each *date query*. In particular, the timestamp is an integer representing the number of days elapsed since 14th January 2008 (the 1st day of the Blogs08 corpus).

For the story ranking task, and in response to a date query, systems should provide a ranking of 100 news stories that they think were important on the specified day (as defined by matching the timestamps between the topic and the news story), for each of the five provided categories of news. When ranking stories, because of the real-time nature of the tackled task, the participating groups were required to only use evidence from blog posts which were published *at or before* the timestamp of the date query, i.e. blog post evidence from after the date query timestamp cannot be used to identify top news.

Figure 2 details the format of a TRC2 story, where the DOCNO tag contains the unique identifier of the story that the system should return; the BLOGS08DAY tag contains the integer timestamp described above; the HEADLINE and CONTENT tags contain the headline and content of the story, as provided by Thomson Reuters.

Figure 3 provides an example of topic illustrating a date query. Only the TRC2 news stories with the same value in the BLOGS08DAY

Group	Run	Baseline	Topic Fields	Facet MAP(baseline run)	Facet MAP(facet run)	Improvement
opinionated						
PKUTM	PKUTM111onB1	PKUTMB1	Q	0.1761	0.2807	59.40%
BIT	BIT10bl2fd3	BITblog10bl2	Q	0.2033	0.2806	38.02%
ULugano	LexMIRuns1	stdbaseline1	Q	0.2598	0.2656	2.23%
HIT_LTRC	hitFeeds1	stdbaseline1	?	0.2598	0.2607	0.35%
factual						
BIT	BIT10bl1fd4	BITblog10bl1	Q	0.2976	0.2987	0.37%
PKUTM	PKUTM211STD1	stdbaseline1	Q	0.2693	0.2761	2.53%
ICTNET	ICTNETFBD1	ICTNETBDRun1	Q	0.2563	0.2740	6.91%
HIT_LTRC	hitTDNfeedR*	N/A	QDN	N/A	0.2735	N/A
official						
BIT	BIT10bl1fd4	BITblog10bl1	Q	0.3312	0.3333	0.63%
ICTNET	ICTNETFBD3	N/A	Q	N/A	0.3134	N/A
PKUTM	PKUTM123STD1	stdbaseline1	Q	0.2439	0.2937	20.42%
HIT_LTRC	hitFeeds1	stdbaseline1	?	0.2439	0.2464	1.03%
personal						
ULugano	LexMIRuns1	stdbaseline1	Q	0.1377	0.2121	54.03%
BIT	BIT10std1fd4	stdbaseline1	Q	0.1377	0.1950	41.61%
PKUTM	PKUTM111onB2	PKUTMB2	QDN	0.1441	0.1901	31.92%
HIT_LTRC	hitTDNfeedR*	N/A	QDN	N/A	0.1549	N/A
indepth						
ICTNET	ICTNETBD4	N/A	Q	N/A	0.3478	N/A
BIT	BIT10bl1fd1	BITblog10bl1	Q	0.2153	0.3211	49.14%
uogTr	uogTrfL728s1	stdbaseline1	Q	0.2345	0.2971	26.70%
PKUTM	PKUTM111onB1	PKUTMB1	Q	0.1644	0.2407	46.41%
shallow						
BIT	BIT10bl1fd4	BITblog10bl1	Q	0.2104	0.2108	0.19%
uogTr	uogTrfC919	uogTrLv450	Q	0.1521	0.1496	-1.64%
UICIR	uicfbdstd2b	stdbaseline2	Q	0.1259	0.1370	8.82%
HIT_LTRC	hitTDNfeedbl*	hitTDNbl	QDN	0.1395	0.1331	-4.59%

Table 8: For each facet, the best faceted blog distillation run from the top four groups sorted by Facet MAP. Facet MAP(baseline run) is the Facet MAP of the baseline ranking on the same facet inclination. * denotes a manual run.

```
<top>
<num>TS10-01</num>
<date>2008-04-24</date>
<day>Wednesday</day>
<blogs08day>100</blogs08day>
</top>
```

Figure 3: Blog track 2010, story ranking task, topic 1 where the num tag contains the topic number and the blogs08day tag contains the integer timestamp described above.

tag as the topic has in the blogs08day tag should be ranked in response to a date query. For example, for a topic with <blogs08day> 5 </blogs08day>, the participating systems should only rank TRC2 news stories with <BLOGS08DAY> 5 </BLOGS08DAY>, using blog post evidence from Blogs08, which have timestamp ≤ 5.

A total of 50 new query dates were randomly sampled from across the timespan of the Blogs08 corpus. The selected dates have a balanced coverage of the months of the Blogs08 collection, as well as the seven days of the week. After the submission of the story ranking task runs, for the purposes of the news blog post ranking task, the organisers selected 68 news stories covering the five categories for which relevant and diverse blog posts should be identified by the participating systems. In particular, for each news story, the participating systems were asked to provide 3 rankings of 50 blog posts, which should be relevant to the news story, and discuss the different aspects of the news stories (e.g. different opinions, type of blog posts, etc). To investigate how blog postings about a story evolve over time, each of the three required rankings is centred at a different period of time:

1. Before the timestamp of the "query date". i.e. blog posts must have timestamp $\leq$ query timestamp
2. One day after the "query date". i.e. blog posts must have timestamp $\leq$ query timestamp + 1 day
3. One week after the timestamp. i.e. blog posts must have timestamp $\leq$ query timestamp + 7 days

3.2 Assessments and Pools

Participating groups were allowed to submit up to 3 runs for the story ranking task. Each run consists of a ranking of 100 news stories for each news category on each query date (i.e. 5 rankings for each given query date). A total of 18 runs were received from 5 groups, including one manual run.

Pools were created using stratified sampling, as defined for the statMAP measure [2]. In particular, 32 news stories for each category and day were sampled from the headlines ranked in the top 30 by any of the submitted runs. In a marked departure from the usually adopted community judgements within TREC, we made a first attempt at using *crowdsourcing* for assessing all the generated pools in this task. In particular, we employed over 720 unique workers from the Amazon's Mechanical Turk to judge 7,427 stories spanning 50 query dates and 5 news categories. Each worker

Group	Approach of	Mean	Mean Facet MAP by Std. Baseline					
			Stdbaseline1		Stdbaseline2		Stdbaseline3	
ULugano	LexMIRuns_	0.1598	0.2180	4.74%	0.1362	-2.52%	0.1250	6.88%
BIT	BIT10_fd1	0.1567	0.2146	3.07%	0.1350	-3.39%	0.1204	2.95%
HIT_LTRC	hitFeeds_	0.1555	0.2089	0.36%	0.1398	0.06%	0.1179	0.76%
UICIR	uicfbd_b	0.1458	0.1588	-23.71%	0.1588	13.67%	0.1197	2.31%
PKUTM	PKUTM123_	0.1404	0.1853	-11.00%	0.1246	-10.80%	0.1112	-4.93%
uogTr	uogTrfL919_	0.1224	0.1837	-11.75%	0.1067	-23.67%	0.0769	-34.24%
UniNE	run3swnpn_0	0.1142	0.1434	-31.10%	0.1016	-27.26%	0.0976	-16.60%
PCUHK	Std_PI	0.0942	0.1006	-51.67%	0.1000	-28.43%	0.0820	-29.89%

Table 9: For each group, the best set of runs for each group, applied over all three standard baselines. In an approach, _ denotes the part of the run name representing the used standard baseline. Some groups did not submit faceted runs using all three baselines.

was shown the pool of news stories for a given category on a given day, and asked to judge each news story as one of the following:

Important and correct category: This is a big story, which should be ranked highly for this category.

Not important but correct category: This story is not very important and should be ranked lower.

Wrong category: This story could be either important or not, but it doesn't matter because it doesn't fit into this category.

To assure quality, we used best practises in crowdsourcing [3, 16], whereby each story was judged by three independent workers, resulting in over 24,000 individual judgments. The majority vote for each story was taken as the final binary relevance label, with *Not important but correct category* and *Wrong category* being collapsed into a single *Not important* label. The labelling resulted in high levels of between-worker agreement, increasing our confidence in the quality of the results. Furthermore, all judgments were subject to a manual validation before being approved. Poor quality and fraudulent judgments were rejected and the work republished for new workers to attempt.

Participating groups were also allowed to submit up to 3 runs for the news blog post ranking task. A total of 11 runs from 4 groups were received - all runs were automatic.

The blog post pools for each of the 68 news stories selected as topics were created from the top 20 blog posts ranked in the preferred run submitted by each group. This resulted in a pool of 7975 blog posts. Each of these blog posts was judged as relevant, possibly relevant or not relevant to the news story they where retrieved for. In summary, each of the 7975 blog posts pooled were judged as to their relevancy to a news story, as shown below.

Relevant: Story is discussed.

Possibly relevant: Post could be discussing the story.

Not Relevant: Story is not discussed.

Furthermore, to enable the assessment of the diversity of each ranking produced, all blog posts were also labelled using a number of predefined perspectives that describe each blog post. In particular, to evaluate the range of perspectives that each blog post ranking covers, for the 68 stories, each blog post was also assigned zero or more of the following nine perspectives:

Factual Account: The post just describes the facts as is.

Opinionated Positive: The post expresses a viewpoint endorsing some aspect of the story.

Relevance Level	# Stories
Not Important	5984
Important	1443

Table 10: Breakdown of relevance levels for the story ranking task judgements.

Opinionated Negative: The post criticises some aspect of the story.

Opinionated Mixed: The post expresses both positive and negative opinions.

Short summary/Quick bites: The post contains only a sentence or two about the story.

Live Blog: The post was continually updated at the time about the story.

In-depth analysis: The post goes into significant detail about the story.

Aftermath: The post gives a round-up or retrospective account of the story.

Predictions: The post was written before the story and discusses what might happen.

The relevance assessment phase resulted in high levels of agreement with a gold standard generated by the track organisers, increasing our confidence in the quality of the results. Further details regarding the crowdsourcing of relevance assessments for these tasks can be found in [10].

3.3 Story Ranking Task Results

In this section, we provide an overview of the the results of the story ranking task, namely the effectiveness of the participating systems in identifying the top news for a given query date. The TRC2 corpus contains 1,613,707 newswire stories published by Thomson-Reuters, an average of 4236 stories per day. After our evaluation, 19% of the pooled stories for each day and category were judged to be important. Table 10 provides the detailed breakdown of the relevance assessment of the pooled stories.

Due to the use of stratified sampling, we report the statMAP evaluation measure [2] for the evaluation of story ranking task runs. Moreover, the TRC2 corpus often has many duplicate stories on a given day, which have been updated with more information as breaking news evolves. To account for these during evaluation, we created equivalence classes of news stories based on headline clustering. Only one news story per equivalence class was judged, and when evaluating runs, only one news story per equivalence class

	Best	Median
Business	0.3155	0.0314
Sci.Tech	0.3009	0.0160
Sport	0.4661	0.0927
U.S.	0.5958	0.1535
World	0.5405	0.0949

Table 11: Best and medians for the story ranking stage of top news stories identification task, broken down by category.

was allowed. This ensured that systems which did or did not perform duplicate removal were treated fairly.

Table 11 reports the best and median statMAP measures for each news category. From this, we note that the U.S. and World categories were the easiest for the systems, perhaps due to their superior coverage in the blogosphere.

Table 12 shows the best submitted run for each group, regardless of run type (automatic or manual) and used TRC2 fields (headline and/or content). They are ranked by the mean statMAP over each of the 5 categories, denoted Mean statMAP. The top-performing run was submitted by POSTECH KLE, and used a probabilistic model that considers events, news stories and blog posts. The ikm100 system used a headline-post network structure to identify important stories. ICTNET treated the headline and content of each news story as a query, and accumulated the BM25 scores for relevant blog posts on each day. The UoS group identified the terms whose frequencies in blog posts increased substantially on the day of the query. These terms where then used as a query to rank the stories, using the Terrier platform and its PL2 weighting model. The uogTr group used a learned voting technique to rank news stories for a day of interest. In particular, the ranking is learned using 1076 voting features, extracted using 8 story representations and varying temporal evidence from the 10 days before the day of interest. Finally, ULugano used a clustering method to identify the most important terms on a given day, which are then used to rank news stories.

3.4 News Blog Post Ranking Task Results

In this section, we provide an overview of the results of the News Blog Post Ranking Task, specifically the effectiveness of participating systems at retrieving blog posts related to a news story in a real-time manner. As noted earlier, the track organisers selected 68 news stories, each comprised of a headline, some article content and a date, which act as the topics that systems were to rank blog posts for. For all 68 news stories, participating systems were to return three distinct rankings, representing searches at three points in time relative to the time the story was published. In particular, for each news story, systems were to retrieve blog posts from:

1. Before the story was published (*Real-time*)
2. One day after the story was published and before (*+1 day*)
3. Seven days after the story was published and before (*+7 days*)

A total of 11 runs from 4 groups were received. A run was evaluated based upon the number, ranking and diversity of relevant blog posts contained. The primary evaluation metric for the news blog post ranking task is α-Normalised Discounted Cumulative Gain at rank 10 (α-nDCG@10). This measure incorporates both support for the three level graded relevance judgments used, and promotes systems that diversify their rankings in terms of the nine perspectives described earlier in Section 3.2.

Table 13 reports the best and median α-nDCG@10 measures over all runs, both in terms of the mean of the three rankings in each run and each type of ranking individually (*Real-time*, *+1 day* and *+7 days*. From this, we note that as time progresses, the effectiveness of systems tends to increase. This is intuitive, as over time, new blog posts discussing each story will be posted.

	Best	Median
all	0.6097	0.4207
Real-time	0.6070	0.4137
+1 Day	0.6044	0.4185
+7 Days	0.6176	0.4298

Table 13: Best and medians α-nDCG@10 for the blog post ranking stage of top news stories identification task, broken down by category.

Table 14 reports the best run submitted by each of the four groups in terms of mean α-nDCG@10 over all three rankings per run in addition to the α-nDCG@10 score for each ranking type individually. Runs are ranked based upon the mean α-nDCG@10 reported. The best performing run was that submitted by uogTr, and leveraged a learning to rank approach over 81 blog post features to rank blog posts for each story. Notably this run did attempt to diversify the blog post rankings. POSTECH KLE also applied an effective diversification strategy, which considers both the relevance and similarity between a news story and blog posts. The ICTNET system employed an ensemble ranking strategy to rank blog posts but did not apply any diversification. The ikm100 system ranked blog posts based upon the normalised cosine similarity between each news story headline and each blog post considered and did not diversify the rankings.

4. CONCLUSIONS

In its fifth year, the TREC Blog track has tackled advanced tasks in the form of faceted blog distillation and top news identification. In both tasks, compared to TREC 2009, sub-tasks have been formulated that allow the effect of components of participants systems to be evaluated independently. It is of note that the relevance assessments of the top stories identification task have been obtained through crowdsourcing, the first successful attempt of its kind within a TREC track.

TREC 2010 represents the final year of the Blog track in its current form. Over the past five years, we have developed test collections for several user search tasks on the blogosphere, namely opinion-finding, blog distillation (aka feed search) and top news identification. Two blog corpora have been developed, namely Blogs06 and Blogs08, and two news corpora (NYT and TRC2) have been released for the benefit of the information retrieval (IR) community. We believe that these corpora and test collections will be valuable to the IR researchers and practitioners for sometime to come. In TREC 2011, the Blog track will morph into the Microblog track, and will tackle search tasks prevalent on micro-blogosphere such as social and real-time search.

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 Thomson Reuters for providing the large sample of headlines and content used in the top stories identification task. They are provided to support research in the TREC blog track. Thanks are also due to Richard McCreadie for his help and assistance with the use of crowdsourcing and for contributions to this overview.

Group	Run	TRC2 Fields	Mean statMAP	statMAP by Category				
				Business	Sci-Tech	Sport	U.S.	World
POSTECH_KLE	KLERUN1	HC	0.2206	0.1851	0.1821	0.1916	0.2458	0.2986
ikm100	ikm100jing	HC	0.2151	0.1144	0.2483	0.1725	0.3897	0.1504
ICTNET	ICTNETTSRun2	HC	0.2138	0.0969	0.1898	0.2405	0.3025	0.2396
UoS	strath2*	HC	0.1285	0.0218	0.0029	0.2308	0.1275	0.2595
uogTr	uogTrLC151	HC	0.1139	0.0907	0.0058	0.1066	0.1230	0.2434
ULugano	CombMNZ	HC	0.1000	0.0428	0.0698	0.0926	0.2801	0.0149

Table 12: Top stories identification task: Ranking of runs for identifying important stories, one run per group. Ranked by Mean statMAP over all categories. * denotes a manual run.

Group	Run	Mean α-nDCG@10	α-nDCG@10 by Category		
			Real-time	+1 Day	+7 Days
uogTr	uogTrL81	0.4771	0.4688	0.4671	0.4953
POSTECH_KLE	KLE1	0.4651	0.4665	0.4626	0.4663
ICTNET	ICTNETPRRun3	0.4266	0.4255	0.4175	0.4368
ikm100	run3	0.4075	0.3779	0.4096	0.4350

Table 14: Top stories identification task: Ranking of runs for blog post ranking, one run per group. Ranked by Mean α-nDCG@10 over all categories.

5. REFERENCES

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

[2] J.A. Aslam and Virgil Pavlu. A Practical Sampling Strategy for Efficient Retrieval Evaluation. Technical Report, North Eastern University.

[3] C. Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of EMNLP 2009*, Singapore.

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

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

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

[7] C. Macdonald, I. Ounis, and I. Soboroff. Overview of TREC-2009 Blog track. In *Proceedings of TREC 2009*, Gaithersburg, USA, 2010.

[8] 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`

[9] C. Macdonald, R.L.T. Santos, I. Ounis and I. Soboroff. Blog track research at TREC. SIGIR Forum, 44(1):58–75, 2010.

[10] R. McCreadie, C. Macdonald and I. Ounis. Crowdsourcing Blog Track Top News Judgments at TREC. In *Proceedings of CSDM 2011*, Hong Kong, China, 2011.

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

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

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

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

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

[16] R. Snow, B. O'Connor, D. Jurafsky and A. Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP 2008*, Honolulu, USA, 2008.

[17] 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)*, Edinburgh, UK, 2006.

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