On the TREC Blog Track

Iadh Ounis, Craig Macdonald and **Ian Soboroff**
University of Glasgow and NIST
UK and USA
{ounis,craigm}@dcs.gla.ac.uk and ian.soboroff@nist.gov

Abstract

The rise of blogging as a new grassroots publishing medium and the many interesting peculiarities that characterise blogs compared to other genres of documents opened up several new interesting research areas in the information retrieval field. The Blog track was introduced in 2006 as part of the renowned Text REtrieval Conference (TREC) evaluation forum, to drive research on the blogosphere and to facilitate experimentation and evaluation of blog search techniques. This paper reports on two years of the Blog track at TREC. We describe the blog search tasks investigated at TREC, and discuss the main lessons we learnt from the track. We conclude the paper with discussions of the broader implications of the Blog track lessons and possible directions for the future, with the aim to uncover and explore the richness of information available in the blogosphere.

Introduction

The growth of interest in blogs, and the richness of information available on the blogosphere has opened up several new interesting research areas in the information retrieval (IR) field. Indeed, the need to have appropriate retrieval techniques to track and find out about the way bloggers react to products, trends and events as they unfold raises some challenging problems in IR. In particular, the problem of retrieving and analysing non-factual aspects of information such as opinions, sentiments, perspectives or personal experiences remains open.

The development of new and appropriate retrieval techniques for the blogosphere requires a suitable infrastructure for experimentation and evaluation along with realistic datasets. The IR community has a long standing history in experimentation and evaluation (Voorhees 2007), as exemplified by the annual Text REtrieval Conference (TREC), an internationally acclaimed forum organised by the National Institute of Standards and Technology (NIST, USA), since 1992. TREC aims to support IR research by providing the infrastructure necessary for large-scale evaluation of text retrieval techniques and approaches. The idea behind TREC is to evaluate IR systems on standard and controlled test collections. A test collection consists of three components: a collection of documents; an associated set of information need statements called "topics"; and a set of human relevance judgements stating which documents are relevant for which topics. A reusable test collection with its associated topics and relevance judgements is very important in encouraging IR research. It allows for the reproducibility of results, and facilitates the further development of new retrieval techniques, the effectiveness of which can be compared to other existing techniques on the same collection.

Given the large size of the collections, the relevance judgements cannot be exhaustive. Instead, TREC uses a process called "pooling" (Spärck Jones & van Rijsbergen 1975), where for each topic the assessors do not judge all documents in the collection, but only the top-ranked documents (usually 100) by the set of the participating search engines. Traditionally, TREC search tasks are described as "adhoc", where the retrieval effectiveness of a system is assessed by its ability to rank, for each topic, as many as possible of the documents judged relevant by the human assessors above the rest of the documents. Adhoc search tasks are usually evaluated using mean average precision (MAP) over all topics. Average precision (AP) for one topic is mean of the precision values calculated as each relevant document is retrieved.

TREC is organised in "tracks", each addressing specific search tasks on a given collection of documents of different genres, ranging from newswire articles to Web pages, email messages and video clips. In TREC 2006, a new Blog track was introduced. It aims to support research in exploring the information seeking behaviour in the blogosphere. The Blog track addressed two main search tasks: Firstly, the opinion finding task addresses a key feature that distinguishes blog contents from the factual content traditionally used in other TREC search tasks, namely the subjective nature of blogs (Mishne & de Rijke 2006); secondly, the blog distillation task is concerned with the search of blogs rather than blog posts.

This paper draws conclusions from two years of the Blog track in TREC 2006 (Ounis *et al.* 2007) and TREC 2007 (Macdonald, Ounis, & Soboroff 2008), assessing the progress made so far, and proposing new research directions. The next section describes the TREC Blog track and its corresponding Blog06 test collection. This is followed by a section that discusses the opinion finding task. We summarise the main effective opinion detection approaches deployed by the participating search engines and provide insights on how well the systems perform across a variety of topic categories. We also report the extent to which spam infiltrates the retrieved blog posts across the topic categories. The following section describes the blog distillation search task. We summarise the main retrieval techniques used for feed search, and also report the performance of the participating search engines and how spam infiltrates the retrieved blogs across a variety of topic categories. The penultimate section discuss the lessons learnt from the first two years of the Blog track, and propose a set of possible interesting future search tasks. Finally, we conclude with the broader implications of the Blog track.

Blog Track at TREC

Similar to all other TREC tracks, the Blog track aims to act as an incubator of new research work, creating the required infrastructure to facilitate research into the blogosphere. Indeed, since its creation, the TREC blog track has aimed to define suitable search tasks on the blogosphere. The grassroots ("non-certified") nature of the blogosphere brings with it a number of challenges. Some are technical, such as spam and widespread duplication. Others are more fundamental such as the fact that subjectivity is a core aspect of the blogs and the blog queries (Mishne & de Rijke 2006). As a consequence, the Blog track has investigated how the subjective nature of blogs and blog queries can be incorporated and exploited in the retrieval context. The blog track addressed the opinion finding task, which is an articulation of a user search task, where the information need could be of an opinion, or perspective-finding nature, rather than fact-finding. The opinion finding task was extended in TREC 2007 with an opinion polarity subtask, where the polarity (or orientation) of the opinions in the retrieved documents must also be returned. Some blog search engines allow users to search for authoritative feeds about a given topic. In TREC 2007, we introduced the Blog distillation task, which evaluates systems on how good they are at finding useful and principle blogs relating to a given topic.

As mentioned in the introduction, an important aspect of the TREC paradigm is the creation of a suitable test collection, where the experiments could be conducted and a system's retrieval performance can be evaluated. As a consequence, the Blog track created the Blogs06[1] test collection. The creation process of the collection and its main features are detailed in (Macdonald & Ounis 2006). The Blogs06 collection represents a large sample crawled from the blogosphere over an eleven week period from the 6th December 2005 until the 21st February 2006. The collection is 148GB in size, with three main components consisting of 38.6GB of XML feeds, 88.8GB of permalink documents (i.e. a single blog post and all its associated comments) and 28.8GB of homepages (i.e. the corresponding blog entry page each time the feed was fetched).

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

Table 1: Statistics of the Blogs06 test collection.

Language	Nbr. Permalinks	Percentage (%)
English	2,794,762	86.9
Spanish	64,350	2.0
French	50,852	1.6
German	18,444	0.6
Italian	10,797	0.3
(other)	76,230	2.4
(unknown)	199,736	6.2

Table 2: Breakdown of language statistics of the Blogs06 collection. The languages labelled *Unknown* correspond almost entirely to Asian languages

Over 100,000 blogs were monitored for the eleven week period, generating 3.2 million permalink documents (posts). The permalink documents are used as a retrieval unit for the opinion finding task and its associated polarity subtask. For the blog distillation task, the feeds are used as the retrieval unit. Table 1 shows the main statistics of the Blogs06 collection. Moreover, in order to ensure that the Blog track experiments are conducted in a realistic and representative setting, the collection also includes a significant portion of spam, non-English documents, and some non-blogs documents such as RSS feeds. About 13% of the permalinks in the Blogs06 collection are non-English. In particular, about 6% of the collection is in Asian languages. Table 2 shows the breakdown of language statistics in the Blogs06 collection. It is of note that only English posts are assessed, posts in any other language are deemed non-relevant. Finally, during the creation of the collection, 17,969 presumed spam blogs (known as splogs) and their corresponding 509,137 blog posts were included in the Blogs06 collection to assess the impact of spam on the retrieval performance in such a controlled setting (Macdonald & Ounis 2006).

In the first pilot run of the Blog track in TREC 2006, it was comprised of the opinion finding task, and an open task which allowed participants the opportunity to influence the determination of a suitable second search task for 2007 on other aspects of blogs besides their opinionated nature. TREC 2007 saw the addition of a new main task and a new subtask, namely the blog distillation task and a polarity subtask respectively, along with a second year of the opinion retrieval task. Table 3 provides an overview of the number of participating groups in the track since its inception.

[1] `http://ir.dcs.gla.ac.uk/test_collections`

Year	Tasks	Participants
2006	Opinion Finding Task	14
	Open Task	5
2007	Opinion Finding Task	20
	Polarity Subtask	11
	Blog Distillation Task	9

Table 3: Tasks run over the first two years of the TREC Blog track.

In the remainder of this paper, we present in details the two main tasks that have ran at the TREC Blog track. We describe the tasks in details, as well as the most effective retrieval approaches that the participating groups have deployed. We provide insights on the performances of search engines across a variety of topic categories, as well as how the topic categories were affected by spam.

Opinion Finding

Many blogs are created by their authors as a mechanism for self-expression encouraged by the freely accessible blog software, communicating their opinions and thoughts on any topic of their choice. A study conducted in (Mishne & de Rijke 2006) shows that many queries received by blog search engines seem to be of an opinion, or perspective-finding nature, rather than fact-finding. The opinion finding task is an articulation of an information need that aims to uncover the public sentiment towards a given target entity such as a product, an organisation or a location. A retrieval engine allowing for an effective opinion finding might naturally be used as a tool for supporting many business-intelligence tasks such as brand monitoring, consumer-generated views and feedback analysis, and more generally media analysis. It can also help users make an informed choice before buying a given product, attending an entertainment event, or taking a holiday in a given location.

Several commercial blog search engines aim to allow users to find out about the opinions and thoughts of other people, who happily share their thoughts on the blogosphere. These thoughts range from anger at some products, politicians or organisations, to good reviews of products or appraisal of cultural events.

In the Blog track, the opinion retrieval task involved locating blog posts that express an opinion about a given target (Ounis *et al.* 2007). The target can be a "traditional" named entity, e.g. a name of a person, location, or organisation, but also a concept (such as a type of technology), a product name, or an event. The task can be summarised as *What do people think about X*, *X* being a target. The topic of the post is not required to be the same as the target. However, for a post to be judged relevant, an opinion about the target had to be present in the post or one of the comments to the post, as identified by the permalink. To create a realistic setting where the topics are actual representations of real information needs, assessors selected queries from a query log of a commercial blog search engine, and expanded them into fully-described topics by making a reasonable interpretation of the query. This process was used to generate 50 topics for the 2006 Blog track and another 50 topics for the 2007 Blog track. Figure 1 shows an example topic.

```
<top>
 <num> Number: 930 </num>
 <title> ikea  </title>
 <desc> Description:
  Find opinions on Ikea or its products.
 </desc>
 <narr> Narrative:
  Recommendations to shop at Ikea are relevant opinions.
  Recommendations of Ikea products are relevant opinions.
  Pictures on an Ikea-related site that are not related
  to the store or its products are not relevant.
 </narr>
</top>
```

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

The relevance assessment procedure had two levels (Ounis *et al.* 2007; Macdonald, Ounis, & Soboroff 2008). The first level assesses whether a given blog post, i.e. a permalink, contains information about the target and is therefore relevant. The second level assesses the opinionated nature of the blog post, if it was deemed relevant in the first assessment level. A workable definition of *subjective* or *opinionated* content was used. In particular, a post is assumed to have a subjective content if it contains an explicit expression of opinion or sentiment about the target, showing a personal attitude of the writer. Rather than attempting to provide a formal definition, the human assessors were given a number of examples, which illustrated the two levels of assessments. Given a topic and a blog post, assessors were asked to judge the content of the blog posts. The following scale was used for the assessment:

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

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

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

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

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

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

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

Following the TREC paradigm described in the previous section, the relevance assessments were formed using the

Relevance Scale	Label	#(2006)	#(2007)
Not Relevant	0	47491	42434
Relevant	1	8361	5187
Negative Opinionated	2	3707	1844
Mixed Opinionated	3	3664	2196
Positive Opinionated	4	4159	2960
(Total)	-	67382	54621

Table 4: Opinion finding task: Relevance assessments of documents in the pool.

pooling technique and the Blogs06 test collection. Table 4 shows the breakdown of the relevance assessments of the pooled documents using the two levels assessment procedure described above for the Blog track 2006 and 2007. In both years, there were slightly more positive opinionated documents than negative or mixed opinionated documents, suggesting that overall, the bloggers had more positive opinions about the topics tackled by the TREC 2006 and 2007 opinion finding topic sets.

Looking into the retrieval techniques deployed by the 14 participants in 2006 and the 20 participants in 2007, we note that most participants indexed the permalink component of the Blogs06 collection and approached the opinion finding task as a two-stage process. In the first stage, documents are ranked using modern and effective document ranking functions such as BM25, language models (LM) and divergence from randomness (DFR) models. It is important to note that the two years running of the Blog track showed that a strongly performing baseline, which aims to find as many relevant documents as possible independently of their degree of opinionated content, is very important in achieving a good opinion finding retrieval performance (Ounis *et al.* 2007; Macdonald, Ounis, & Soboroff 2008).

In the second stage of the retrieval process, the retrieved documents are re-ranked taking into account opinion finding features, often through a combination of the first stage retrieval score with a computed score denoting the degree of opinionated content in the document. Results from two years of the opinion finding task show that there are two main effective approaches for detecting opinionated documents. The first approach consists in automatically building a weighted dictionary from a training dataset where the distribution of terms in relevant and opinionated documents is compared to their distribution in relevant but not necessarily opinionated documents. The resulting weight of each term in the dictionary estimates its opinionated discriminability. The weighted dictionary is then submitted as a query to generate a score predicting how opinionated each document of the collection is (in order to speed up retrieval this prediction score can be done at indexing time). The second approach uses a pre-compiled list of subjective terms and opinion indicators and re-ranks the documents based on the proximity of the query terms to the aforementioned pre-compiled list of terms. These two approaches, despite their relative simplicity and very low computational overheads, led to a marked improvement in the MAP of the opinion finding performance, compared to a corresponding baseline where the opinion finding features have been turned off.

Group	Techniques
UIllinois	Concept-based retrieval; Query expansion; SVM classifier
UGlasgow	Field-based DFR document ranking (PL2F); Proximity search; Dictionary-based approach or OpinionFinder;
UArkansas	Language Models; Proximity-based approach
UWaterloo	BM25; Proximity-based approach
FIU	DFR document ranking (DPH); Dictionary-based approach

Table 5: Techniques applied by various groups to the opinion finding task. Group names are of participating groups of TREC Blog track (Ounis *et al.* 2007; Macdonald, Ounis, & Soboroff 2008)

There were two other approaches of interest: An approach based on building a Support Vector Machines (SVM) classifier to estimate the degree of opinionated content in the retrieved documents. This method requires a very extensive training phase based on sources known to contain opinionated content (such as web sites specialising in product reviews) and sources assumed to contain little opinionated content (such as online encyclopedias or news collections). Another interesting approach was based on OpinionFinder, a freely available natural language processing toolkit, which identifies subjective sentences in text (Wilson *et al.* 2005). For a given document, the OpinionFinder tool is adapted to produce an opinion score for each document, based on the identified opinionated sentences. While both these techniques are effective, they have very high computational and/or training overheads. A detailed overview of the techniques deployed by the participating groups in the opinion finding task could be found in (Macdonald, Ounis, & Soboroff 2008; Ounis *et al.* 2007). Table 5 highlights the techniques applied by the best performing groups.

The 100 topics used in TREC 2006 and 2007 address a wide range of categories and target entity types. We manually grouped the 100 topics by the category that they addressed (e.g. travel, music, politics, health), and by their target entity type (e.g. person, product, organisation). The distribution of category and entity types is shown in Table 6.

To assess whether there was a variation in the difficulty of the categories, we perform an analysis based on how well a fictional system would perform for each category - this fictional system is a system that performs the median level of average precision (AP) for each topic out of all participating systems (the use of the median prevents bias towards outliers). Figure 2(a) shows the median AP for each topic, averaged across all topics in each category. While the shopping category looks to be easier than others, and the environment and gaming topics the most difficult, all three categories only contained one topic. Indeed, most categories had a similar difficulty level, in the range from 0.15 to 0.23.

Similarly, Figure 2(b) examines the median AP for each topic, averaged across all topics of each entity type. The figure shows that while topics related to persons, locations and products seemed to be easier to handle by the participating search engines than topics about TV programs, overall, there were no large variations of difficulty in the topics across all entities.

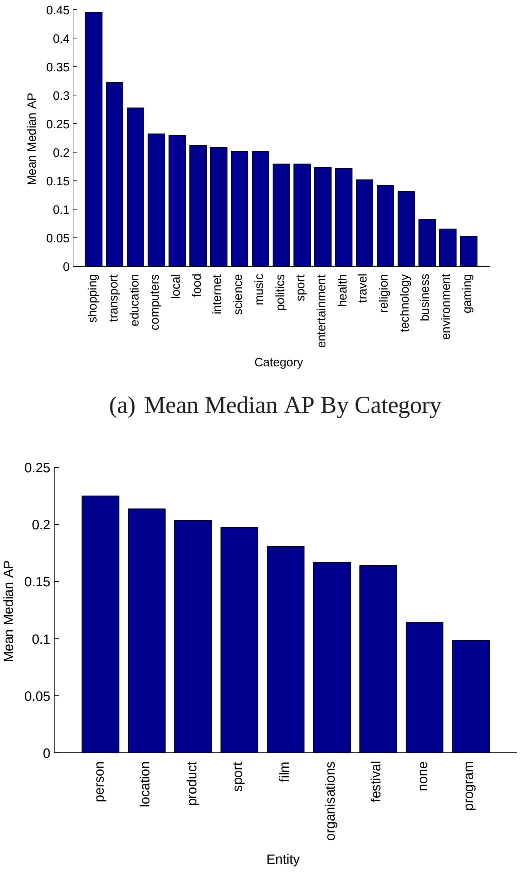
Categories			
politics	23	music	3
entertainment	19	computers	2
sport	9	education	2
food	7	local	2
technology	7	environment	1
internet	6	fashion	1
religion	5	gaming	1
business	4	shopping	1
science	4	transport	1
health	3	travel	1
Entities			
Person	24	Program	9
Organisation	22	Location	6
Product	16	Film	2
Festival/Meeting	10	(none)	10
Sport	9		

Table 6: Distribution of the 100 opinion finding task topics (TREC 2006 and 2007) over categories & entities. (none) denotes when no entity is described.

We examined how the topics and their associated categories and entities have been affected by spam posts. Figure 3(a) shows the extent to which the 17,958 presumed splog feeds and their associated 509,137 posts, which were injected into the Blogs06 collection during its creation, have infiltrated the retrieved documents by the participating search engines per category. We measure the median number of retrieved spam posts retrieved for each topic[2], averaged across all topics in each category. Noticeably, the category with the most splog posts is health. This confirms an analysis of the topics in the Blog track 2006, where systems returned more splog posts for health-related topics (Ounis *et al.* 2007). Retrieved results for the transport, and gaming categories contained more splog posts than the rest of categories (although there was only one topic for each of these two categories). On the other hand, the categories related to technical topics such Internet and science were less infiltrated by splog posts.

Similar to Figure 3(a), Figure 3(b) examines how the median search engine was more likely to retrieve assumed spam blog posts per entity. The figure shows that topics related to sport were more likely to retrieve splog posts. In contrast, the topics related to the person and location entities are the least likely to retrieve splog posts. The fact that topics related to people seem to result in far fewer spam documents in the retrieved documents of the participating search engines has been also noted in (Ounis *et al.* 2007; Macdonald, Ounis, & Soboroff 2008).

Such analysis on a category or entity grouping basis allows blog search engines to identify the areas spammers are targeting, and moreover, to apply selective techniques to deal with particularly easy or difficult categories. This has obvious links to other IR research areas, such as Web query-type classification (Beitzel *et al.* 2005), where different IR techniques may be deployed selectively on a query-by-query basis.



(a) Mean Median AP By Category

(b) Mean Median AP By Entity

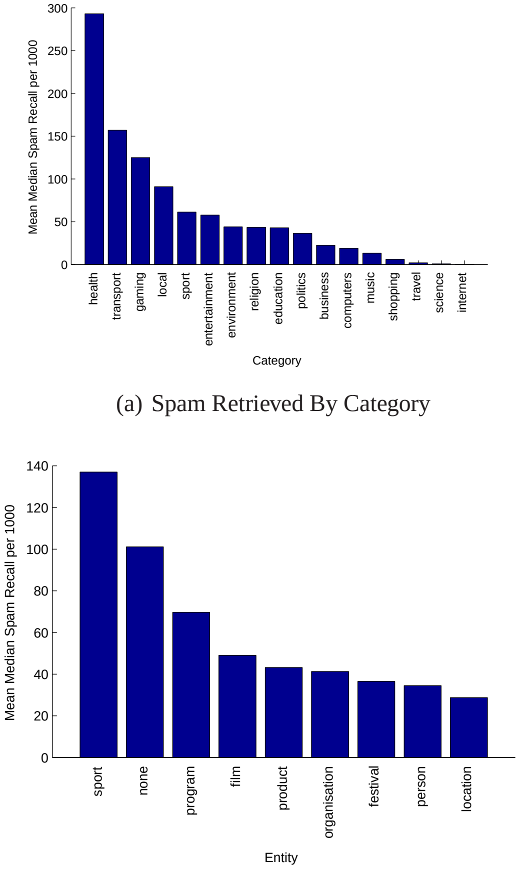
Figure 2: Opinion finding task: Median AP per topic, averaged across all topics in each category/entity.

Blog Distillation

Blog search users often wish to identify blogs about a given topic area, which they can subscribe to and read on a regular basis (Mishne & de Rijke 2006). This user task is most often manifested in two scenarios: **Filtering:** The user subscribes to a repeating search in their RSS reader, looking for new posts about a topic; **Distillation:** The user searches for blogs with a recurring central interest in a given topic, and then adds these to their RSS reader.

In 2007, the TREC Blog track investigated the latter scenario, i.e. a blog distillation task. The blog distillation task can be summarised as *Find me a blog with a principle, recurring interest in X*. For a given area X, systems should suggest feeds that are principally devoted to X over the timespan of the feed, and would be recommended to subscribe to as an interesting feed about X (i.e. a user may be interested in adding it to their RSS reader). While the underlying user task is of an adhoc nature - this task is intended to have a

[2] The spam retrieved is the number of spam blog posts retrieved in the top 1000 returned documents by a search engine for a topic.



(a) Spam Retrieved By Category

(b) Spam Retrieved By Entity

Figure 3: Opinion finding task: Spam Retrieved (out of 1000), averaged across all topics for each category, and each entity.

```
<top>
  <num> Number: 994 </num>
  <title> formula f1 </title>
  <desc> Description:
   Blogs with interest in the formula one (f1) motor
   racing, perhaps with driver news, team news, or
   event news.
  </desc>
  <narr> Narrative:
   Relevant blogs will contain news and analysis
   from the Formula f1 motor racing circuit. Blogs
   with documents not in English are not relevant.
  </narr>
 </top>
```

Figure 4: Blog track 2007, blog distillation task, topic 994.

higher relevance threshold, in that relevant blogs must have a principle and recurring interest in the topic area - i.e. the relevance assessments are distilled from those that would contain all blogs with one or more relevant posts. Indeed, various commercial blog search engines provide a blog search facility in addition to their blog post search facility.

Furthermore, the prevalence of blog directories suggest that there is not yet a suitable way to identify blogs with a recurring interest in a topic area. Many blogs include a blogroll which lists blogs that the blogger reads or thinks are related to their own blogging interests. Recently the rise of tagging/social bookmarking - allowing blog readers to attach additional metadata to describe posts - has done much to improve the situation, but the automatic selection of a few blogs which are primarily blogging about a specific topic area over a time period remains an open problem.

The experimental setup of the blog distillation task at TREC 2007 was as follows (Macdonald, Ounis, & Soboroff 2008):

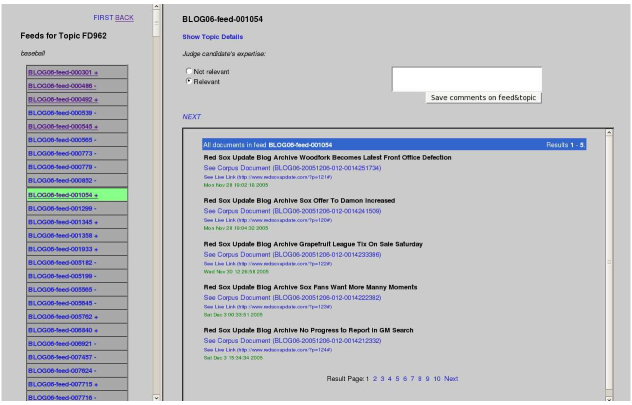
1. From the collection, participants were asked to provide queries (built up into a TREC topic format describing the information need fully), along with a few blogs that met that information need. In all, 45 topics were created by nine participating groups. A sample topic is given in Figure 4.

2. All 45 topics were distributed to participating groups. Groups used their search engines to rank the blogs in the collection in response to the query. These result files were then submitted to TREC for pooling.

3. From the submitted result files from all groups, a pool of possible relevant blogs for each topic were identified by taking the top 50 results from the run of each group. These pools were then used for assessment.

4. Participating groups were asked to assess the relevance of each pooled blog for the topics they created, using a Web-based assessment system. For each blog, the assessment system listed the titles and dates of the posts. Clicking on the title displayed the content of the post. After reading as many or as few as they wish, they make an informed choice of the relevance of the blog, i.e. whether the blog is principally devoted to the topic and would be recommended to subscribe to as an interesting blog about the topic area i.e. a user may be interested in adding the blog's feed to their RSS reader (Macdonald, Ounis, & Soboroff 2008). Principally and recurrent were not explicitly defined, but instead it was suggested that for a blog to be relevant to the topic, a large majority of posts had to be on-topic, over the entire timespan of the blog. Figure 5 provides a screenshot of the assessment system.

From the nine participating groups in the blog distillation task, several approaches for tackling the problem were proposed. Table 7 highlights the best techniques applied by participants. Of particular interest is the work of Carnegie Mellon University (CMU), who examined whether retrieval based on feeds was sufficient in this task, or whether document posts had to be indexed for good retrieval effectiveness. Their experimental results suggest that feeds are indeed sufficient. For those who indexed the posts (permalinks) alone, their techniques worked by combining the scores of all posts of a blog to get a final relevance score per blog. Various groups drew connections to other areas of IR, for instance Distributed IR (CMU & UMass) or Expert Search (UGlasgow), and adapted techniques from these areas to use in their systems.



FIRST BACK

Feeds for Topic FD962

baseball

BLOG06-feed-000301 +
BLOG06-feed-000486 -
BLOG06-feed-000492 +
BLOG06-feed-000539 -
BLOG06-feed-000545 +
BLOG06-feed-000565 -
BLOG06-feed-000773 -
BLOG06-feed-000779 -
BLOG06-feed-000852 -
BLOG06-feed-001054 +
BLOG06-feed-001299 -
BLOG06-feed-001345 +
BLOG06-feed-001358 +
BLOG06-feed-001933 +
BLOG06-feed-005182 -
BLOG06-feed-005199 -
BLOG06-feed-005565 -
BLOG06-feed-005645 -
BLOG06-feed-005762 +
BLOG06-feed-006840 +
BLOG06-feed-006921 -
BLOG06-feed-007457 -
BLOG06-feed-007624 -
BLOG06-feed-007715 +
BLOG06-feed-007716 -

BLOG06-feed-001054

Show Topic Details

Judge candidate's expertise:

○ Not relevant
● Relevant

Save comments on feed&topic

NEXT

All documents in feed **BLOG06-feed-001054** Results **1 - 5.**

Red Sox Update Blog Archive Woodfork Becomes Latest Front Office Defection
See Corpus Document (BLOG06-20051206-012-0014251734)
See Live Link (http://www.redsoxupdate.com/?p=121#)
Mon Nov 28 19:02:16 2005

Red Sox Update Blog Archive Sox Offer To Damon Increased
See Corpus Document (BLOG06-20051206-012-0014241509)
See Live Link (http://www.redsoxupdate.com/?p=120#)
Mon Nov 28 19:04:32 2005

Red Sox Update Blog Archive Grapefruit League Tix On Sale Saturday
See Corpus Document (BLOG06-20051206-012-0014233386)
See Live Link (http://www.redsoxupdate.com/?p=122#)
Wed Nov 30 12:26:58 2005

Red Sox Update Blog Archive Sox Fans Want More Manny Moments
See Corpus Document (BLOG06-20051206-012-0014222382)
See Live Link (http://www.redsoxupdate.com/?p=123#)
Sat Dec 3 00:33:51 2005

Red Sox Update Blog Archive No Progress to Report in GM Search
See Corpus Document (BLOG06-20051206-012-0014212332)
See Live Link (http://www.redsoxupdate.com/?p=124#)
Sat Dec 3 15:34:34 2005

Result Page: 1 2 3 4 5 6 7 8 9 10 Next

Figure 5: Judging in progress for topic 962 ('baseball') of the Blog Distillation task (TREC 2007). The left hand column displays the blogs to be judged. The right hand working area shows the titles and dates of posts from the current blog. Once they have read some of the blog's posts, the assessor can select whether they judge the blog to have a relevant recurring interest in the topic or not. The assessor can save a comment about the feed and topic in the top right text area.

Group	Indexed	Retrieval Technique
CMU	Both	Adaptation of Distributed IR & LM techniques, on either feeds or documents.
UGlasgow	Posts	Adaptation of techniques developed for Expert Search, on top of DFR PL2F.
UMass	Posts	LM & Resource selection (from Distributed IR). Shallow blog penalisation.
UvAmsterdam	Posts	Language modelling approach, combined with the ratio of relevant posts per feed.

Table 7: Techniques applied by various groups participating in the blog distillation task, TREC 2007. Group names are of participating groups of TREC Blog track (Ounis *et al.* 2007; Macdonald, Ounis, & Soboroff 2008)

On a closer examination of the relevance assessments, it seems that some topics generated many relevant blogs. For instance, the assessor for topic 978 ('music') found 153 relevant blogs. In particular, Table 8 shows the distribution of relevant blog ranges, in terms of number of topics. We suggest that for all topics (especially for broad topics), the relevance guidelines should somehow contain an element of importance - i.e. the identification of blogs that have a principle, recurring interest in the topic area, but also that the best blogs (Java *et al.* 2007).

# Rel Feeds	≤ 10	11-20	21-30	31-40	41-90	≥ 91
# Topics	6	6	12	4	8	9

Table 8: Distribution of topics with various number of relevant blogs in the TREC 2007 blog distillation task.

technology	6	computers	2
politics	5	health	2
business	4	internet	2
environment	4	music	2
food	4	gaming	1
entertainment	3	travel	1
science	3	local	1
sport	3	shopping	1

Table 9: Categories in the Blog Distillation task topics, TREC 2007.

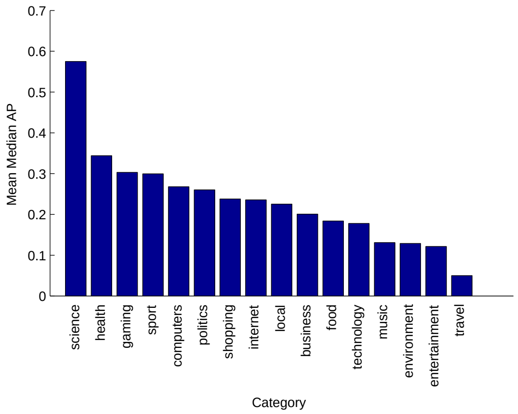


Figure 6: Blog distillation task: Median AP per topic, averaged across all topics in each category.

Similar to the analysis performed for the opinion finding task above, we also manually classified the 45 blog distillation task topics by category of interest. Table 9 presents the distribution of the topics over the categories. We attempted to classify the query by entity type, but found that in contrast to the opinion-finding task topics, the blog distillation task topics were not concerned with entities, but with more general concepts. Hence, the entity analysis was discarded.

Similar to the analysis in the previous section, Figure 6 examines the mean median AP for each topic, averaged across all topics in each category. From the figure, we can draw a few conclusions, namely: that most participating search engines found the three science topics easier than the other topics; most categories were comparatively of medium difficulty (in the range 0.18 - 0.33); topics on music, environment, entertainment and travel were more difficult.

Figure 7 examines how the median system was more likely to retrieve the assumed 17,969 splogs per category. Noticeable from the figure is that the topics from the travel category were most likely to have retrieved splogs (however, with one topic, this could be an outlier). Similarly, the health and shopping topics had more splogs being retrieved than topics in other categories. Contrasting with the opinion finding task (Figure 2(a)), we note that the science topic is much more affected by spam in the blog distillation task systems. As for the opinion finding task, if more topics were available for the sparse categories, the future selective application of retrieval techniques would be an interesting study: e.g. applying spam detection techniques on topic categories more likely to be infiltrated by splogs.

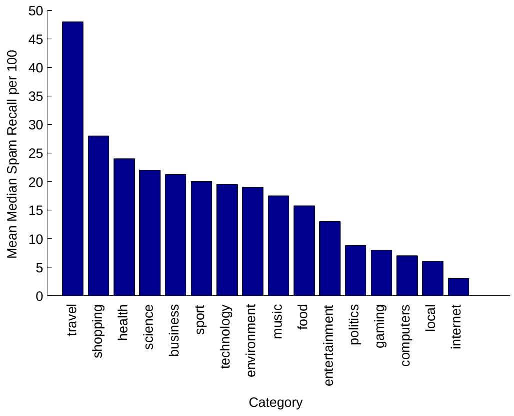


Figure 7: Blog distillation task: Spam Retrieved (out of 100), averaged across all topics for each entity.

To conclude, we believe that the blog distillation task ran successfully during its first year (TREC 2007). Further tightening up of the guidelines should channel this task into a true distillation task that not only finds principle and recurring blogs but also the important ones for general topics.

Lessons Learnt and Future Tasks

In this section, we draw conclusions and lessons learnt from the first two years of the TREC Blog track and provide possible directions for the future. We propose new search tasks for the blogosphere aiming to uncover and explore the richness of information contained in the blogs.

One of the lessons learnt from the opinion finding task is that a good performance in opinion finding is strongly dominated by its underlying document ranking performance (topic-relevance), i.e. the ability of the search engine to retrieve as many relevant documents as possible independently of the degree of their opinionated content (Ounis *et al.* 2007; Macdonald, Ounis, & Soboroff 2008). This suggests that the core retrieval technologies developed by the IR community are still robust and appropriate on this new search task. Another lesson learnt is that simple and light-weight opinion finding detection techniques such as the dictionary-based or the proximity-based approaches are very effective, markedly improving a strong topic-relevance baseline where all opinion finding features were turned off (Macdonald, Ounis, & Soboroff 2008). Note that while SVM-based classifiers can also be very effective in opinion finding, the huge efforts required in training and the possible differences between the training data and the actual opinionated content in the blog posts might make such an approach less practical.

The effectiveness of the dictionary-based approach in opinion finding, coupled with the need of a strong underlying document ranking system has raised the natural question whether the blog track should move on from opinion detection, since enough conclusive and sufficient lessons have been learnt about this task. One area of future direction for the opinion finding task is to move towards opinion retrieval and extraction at the sentence or passage level rather than at the document level. For example, for a given product, one might wish to have a list of its positive and negative features, supported by a set of opinionated sentences extracted from blogs (Popescu & Etzioni 2005). Such a task complements work in the TREC Question Answering track.

The blog distillation task has generated some very promising and interesting retrieval techniques. Interestingly, the techniques that worked best for the blog distillation task are different from those that led to the best retrieval performance on the opinion finding task. Indeed, the approaches that worked best for the former are a refinement and adaptation of IR techniques usually deployed in distributed search or expert finding tasks. In addition, while the feeds component of the Blog06 collection has barely been used in the opinion finding task, some participating groups (e.g. CMU) have shown that it is sufficient for the blog distillation task. This highlights the differences between the two tasks and the need for different approaches for them. As discussed above, the blog distillation task can be further refined by requiring the participating search engines to return the best, instead of all, blogs with a principle and recurring interest for a given topic, similar to the *Feeds that matter* search task described in (Java *et al.* 2007). We would then assess whether techniques that measure authority or popularity can bring retrieval performance benefit.

Splogs are usually recognised as a severe problem on the blogosphere. The Blog track experience is that for for the opinion finding task, there appears to be no strong evidence that spam was a major hindrance to the retrieval performance of the participating search engines. This is possibly because some assumed splogs contain posts with real opinions about the target, which were (rightly) judged as opinionated and relevant by assessors. For the blog distillation task, participants found spam detection techniques to be helpful, however it was not necessary to remove splogs to achieve an effective retrieval approach (Macdonald, Ounis, & Soboroff 2008; Ounis *et al.* 2007). It has been reported that 20% of blogs and 50-75% of pings are splogs (Kolari 2007). The 18% of splogs in the Blog06 collection would therefore seem to be reasonable, so we remain open-minded as to why spam has not severly impacted the Blog track tasks.

The future of the Blog track and the development of new search tasks have been the object of discussion at the TREC 2007 conference. Below, we highlight some of the generic search tasks as well as some variants, which might address some interesting features of the blogosphere:

Feed/Information Filtering: The concept of filtering can be applied to blogs, as a user may express an interest in a topic, and wishes to be informed of new entries that are relevant to his/her desired topic. Indeed, many commercial blog search engines offer such a service, where users subscribe to the search results of a given query, so as to be informed of new relevant posts. A study of commercial query logs in (Mishne & de Rijke 2006) confirmed the prevalence of filtering queries submitted to a blog search engine. They found that a large number of queries are of a repetitive nature, i.e. caused by automatic searches by end-user tools to identify new and timely articles about a general interest area.

The aim of running a filtering task would be to develop techniques for identifying new and timely relevant blog sources. Such a task complements the blog distillation task. The task can be summarised as *Inform me of new feeds or new blog posts about X*. This task might also be tackled from a social-media perspective. For example, users might wish to track posts or comments from a specific person, or related to a particular discussion of a topic across multiple feeds. (i.e. find me blog posts that discuss so-and-so's opinion of the iPod).

Story Detection: The query logs from the commercial search engines show that there is a fair number of news-related queries (Mishne & de Rijke 2006), suggesting that blog search users have an interest in the blogosphere response to news stories as they develop. A possible interesting task would be to have a story tracking task, where, given a starting story or topic, the system should identify all the posts that discuss it. The task can be run with the additional constraint that time is taken into account (Qamra, Tseng, & Chang 2006). This task can be summarised as *Identify all posts related to story X*. A possible variant is to ask the participating systems to provide the *top important* stories or events for a given date or a given range of dates. Conversely, a related timelining event task could also be investigated. It consists in identifying the date of a known event, using evidence from the corpus, and ordering the events in chronological order.

Information Leaders: These are the blogs that break novel information first (Song *et al.* 2007). Given that these blogs are the first to disseminate a story, a new information or an opinion, it is often these blogs that people link to. As a consequence, they can be seen as representative and influential blogs. This task can be summarised as *Identify all information leaders about topic X*, and addresses the social networks/propagation issues in the blog setting.

Overall, there are a variety of possible search tasks that could be investigated as part of the Blog track in the future. One aspect we would like to explore is the substantial expansion of the currently used Blog06 collection both in terms of timespan and number of blogs and posts in the collection. Increasing the timespan of the collection to say 9 or 12 months would allow a more in-depth study of the chronological evidence in the blogosphere.

Conclusions

At its core, TREC is an incubator and a facilitator for research in emerging IR topics and related areas. The Blog track, and other current tracks, are manifestations of this. In its first two years at TREC, the Blog track has investigated two user tasks, lessons have been learnt and conclusions drawn. TREC is an inclusive process: it requires input from many participating groups to successfully run a track, while the outcome is beneficial for the greater IR and related community, by the creation of a reusable test collection.

The research outcomes from the TREC Blog track have a broader interest than to the core IR community. Indeed, researchers from social-media, natural language processing and data mining can have an interest in the research outcomes from the Blog track. Their input to define future tasks, and help in shaping the future of the Blog track is important.

Acknowledgements

Thanks to Maarten de Rijke and Gilad Mishne for their assistance in co-ordinating the TREC 2006 Blog track. Thanks also to the participants of the TREC 2006 & 2007 Blog tracks.

References

Beitzel, S. M.; Jensen, E. C.; Frieder, O.; Grossman, D.; Lewis, D. D.; Chowdhury, A.; and Kolcz, A. 2005. Automatic web query classification using labeled and unlabeled training data. In *Proceedings of SIGIR 2005*, 581–582.

Java, A.; Kolari, P.; Finin, T.; Joshi, A.; and Oates, T. 2007. Feeds That Matter: A Study of Bloglines Subscriptions. In *Proceedings of ICWSM 2007*.

Kolari, P. 2007. *Detecting Spam Blogs: An Adaptive Online Approach*. Ph.D. Dissertation, University of Maryland Baltimore County.

Macdonald, C., and Ounis, I. 2006. The TREC Blogs06 collection : Creating and analysing a blog test collection. Technical Report TR-2006-224, University of Glasgow.

Macdonald, C.; Ounis, I.; and Soboroff, I. 2008. Overview of the TREC-2007 Blog Track. In *Proceedings of TREC 2007)*.

Mishne, G., and de Rijke, M. 2006. A study of blog search. In *Proceedings of ECIR 2006*, 289–301.

Ounis, I.; de Rijke, M.; Macdonald, C.; Mishne, G.; and Soboroff, I. 2007. Overview of the TREC 2006 Blog Track. In *Proceedings of TREC 2006)*.

Popescu, A.-M., and Etzioni, O. 2005. Extracting product features and opinions from reviews. In *Proceedings of HLT 2005*.

Qamra, A.; Tseng, B.; and Chang, E. Y. 2006. Mining blog stories using community-based and temporal clustering. In *Proceedings of CIKM 2006*, 58–67.

Song, X.; Chi, Y.; Hino, K.; and Tseng, B. 2007. Identifying opinion leaders in the blogosphere. In *Proceedings of CIKM 2007*, 971–974.

Spärck Jones, K., and van Rijsbergen, C. 1975. Report on the need for and provision of an 'ideal' information retrieval test collection. British Library Research and Development Report 5266, University of Cambridge.

Voorhees, E. M. 2007. TREC: Continuing information retrieval's tradition of experimentation. *Commun. ACM* 50(11):51–54.

Wilson, T.; Hoffmann, P.; Somasundaran, S.; Kessler, J.; Wiebe, J.; Choi, Y.; Cardie, C.; Riloff, E.; and Patwardhan, S. 2005. Opinionfinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005*, 34–35.