Blog Track Research at TREC

Craig Macdonald, Rodrygo L.T. Santos, Iadh Ounis
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ, UK
{*craigm,rodrygo,ounis*}*@dcs.gla.ac.uk*

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

Abstract

The TREC Blog track aims to explore information seeking behaviour in the blogosphere, by building reusable test collections for blog-related search tasks. Since, its advent in TREC 2006, the Blog track has led to much research in this growing field, and encapsulated cross-pollination from natural language processing research. This paper recaps on the tasks addressed at the TREC Blog track thus far, covering the period 2006 - 2009. In particular, we describe the used corpora, the tasks addressed within the track, and the resulting published research.

1 Introduction

User generated content (UGC) has been a major aspect of the Web 2.0 era. In particular, the prevalence of user-friendly tools has allowed average persons to easily create and publish content online. Often, such UGC takes the form of a blog—a chronologically arranged journal. Recent estimates put the size of the blogosphere (all online blogs) at over 133 million [56].

The Text REtrieval Conference (TREC) is an on-going forum organised by NIST since 1992 to facilitate empirical research into information retrieval (IR) tasks [48, 53]. This research is organised into tracks, each focused around building various test collections, whereby retrieval systems are evaluated on their ability to identify relevant documents in response to a set of test queries. Different tracks have different focuses for their test collections, e.g. newswire articles, Web pages or email messages.

The Blog track at TREC started in 2006, with the aim to explore the information seeking behaviour in the blogosphere. Various tasks have been thus far investigated by the Blog track, namely opinion-finding, blog distillation and top news identification. To enable these tasks, we—the track organisers—have been responsible for creating resources, defining and managing search tasks, and evaluating the approaches of TREC participants.

The paper describes the tackled search tasks within the Blog track thus far, and provides a comprehensive survey of the main published approaches to deploying effective solutions to these tasks. Section 2 describes the two corpora that have been used by the Blog track, namely Blogs06 and Blogs08. Sections 3, 4 and 5 describe each Blog track task in turn, namely opinion-finding, blog distillation and top news, respectively. For each task, we describe its methodology, as well as summarise the successful approaches published in the literature. Finally, in Section 6, we summarise the first four years of the Blog track, provide pointers to further reading, and establish future directions for the track. The review contained in this paper demonstrates the success of the Blog track in facilitating and fostering research in this growing field.

2 Blog Corpora

A corpus is a fundamental component of developing a test collection for IR research. For the Blog track, a large common sample of the blogosphere was required, which could be used by all users of the created test collections. A *blog* represents a chronological ordering of *blog posts* written by one or a few *bloggers*. Over time, other readers of a blog may add *comments* to each blog post. For a usable

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

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

corpora, we required a large sample of blog posts collected over a substantial time period, representing many different blogs. For the TREC 2006 - 2008 campaigns, we created the Blogs06 corpus. This corpus is described in further detail in Section 2.1. The larger Blogs08 corpus was introduced for TREC 2009, and is detailed in Section 2.2.

2.1 The Blogs06 Corpus

The Blogs06 corpus is a sample of the blogosphere crawled over an eleven week period from December 6, 2005 until February 21, 2006. The collection is 148GB in size, with three main components consisting of 38.6GB of XML feeds (i.e. the blog), 88.8GB of permalink documents (i.e. a single blog post and all its associated comments) and 28.8GB of HTML homepages (i.e. the main entry to the blog). In order to ensure that the Blog track experiments are conducted in a realistic and representative setting, the collection also includes spam, as well as some non-English documents.

The Blogs06 corpus was created in several stages:

- Firstly, the set of RSS/Atom feeds to monitor was decided. These included some assumed splog feeds, as well as known "top-blogs" provided by a commercial company. We also endeavoured to ensure that there were blogs of interest to a general audience, by targeting health, travel and political blogs, in addition to the usual personal and technology blogs. In total, over 100,000 blogs were identified, each by an RSS or Atom XML feed.
- Next, the feed list was split into 7 parts, to ensure that we were not accessing large blog hosting providers (e.g. Blogspot.com, Wordpress, etc) too frequently while crawling feeds. Every day, each feed in the feed set for that day was downloaded, as well as the corresponding blog homepage. The links to blog post permalinks found in the downloaded feeds were recorded.
- After a delay of no less than 2 weeks, batches of permalinks (the full content of blog posts with comments) were downloaded. The two week delay was added such that a given blog post may have garnered some comments.
- Finally, after the end of the crawl period, the documents were numbered and ordered by time, to suit the purposes of a TREC test collection.

The finished corpus has a total size of 148GB. The number of permalink documents in the collection amounting to over 3.2 million, while the number of feeds is over 100,000 blogs. Further information on the collection and how it was created can be found in [26]. Some salient statistics of the Blogs06 corpora are listed in Table 1. The Blogs06 corpus was used for the Blog track in years 2006–2008.

2.2 Blogs08 Corpus

After TREC 2007, it became clear that a new and larger blog corpus would be needed to allow tasks with temporal aspects in future Blog track campaigns. To this end, we started the creation of the

Blogs08 corpus in late 2007. In particular, the desired properties were more feeds, more blog posts, collected over a significantly longer time period than the 11 weeks of Blogs06. Firstly, in addition to the feeds in Blogs06, we collected more feeds by sampling from online blog directories, from the recent updates list of major blog hosting providers (e.g. Blogspot and Wordpress). Similarly to Blogs06, we also used blog search engines, to search for blogs of interest to a general audience. Finally, we used outgoing links from blogs in Blogs06 to identify further blogs. Over 1 million blogs were identified using these processes. Note that, in contrast to Blogs06, we did not add any particular spam blog feeds to Blogs08—however, it is highly likely that the collection does contain some.

The crawling strategy for Blogs08 is almost identical to that used for Blogs06. One small difference was that blog homepages were only collected once, rather than each week. We monitored the 1 million blogs on a weekly basis from 14th January, 2008 to 10th February, 2009. This timespan of over 1 year allowed a substantial sample of the blogosphere to be obtained and facilitates studying the structure, properties and evolution of the blogosphere, as well as how the blogosphere responds to events as they happen. Moreover, this time period covered a full US election cycle. While the need for Blogs08 was identified during TREC 2007, we continued to use Blogs06 for TREC 2008 to permit a large number of topics for the tasks using this corpus. Moreover, this allowed the Blog track to continue over the large crawling timespan of Blogs08. Hence, Blogs08 was first used for TREC 2009. Salient statistics of the Blogs08 corpus are also included in Table 1. Both Blogs06 and Blogs08 are distributed by the University of Glasgow[1].

3 Opinion-Finding Tasks

The blogosphere responds to real-world events, in that bloggers author blog posts, discoursing their thoughts on topics of interest. Some topics may 'flow' through the blogosphere, as more bloggers read and re-post their thoughts on a topic [12]. Other topics may become the subject of intense debate among several bloggers in a community, before fading away [21].

This subjective nature of the blogosphere has made it useful to those interested in determining the opinion trends about a topic[2]. For instance, this can provide companies with marketing data about the positive or negative 'buzz' of the blogosphere about various topics. Similarly, a study of a query log from a commercial blog search engine found that many blog search queries seem to be related to uncovering public opinions about a given target [36].

3.1 Task Definition

With the above motivations in mind, we devised the opinion-finding task as part of TREC 2006. In this task, the aim for each query is to identify blog posts expressing an opinion about a given target. 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. An example topic is shown in Figure 1 below.

Notice that there are two aspects to this task. Firstly, *relevance*, whereby the blog posts must be relevant to the target entity in the query topic. Secondly, *opinionatedness*, whereby the blog post must express an opinion about the target entity. In both cases, the task is a ranking problem, and hence systems should rank blog posts that they predict to be relevant and opinionated.

The opinion-finding task ran as part of the TREC Blog track from 2006 to 2008, using the Blogs06 corpus of blog posts. Table 2 details the topic numbers for the total of 150 topics available for this task. Relevance assessments are at the blog post level, and detail whether the blog post was relevant, and if so, what kind of opinion was expressed towards the target entity (positive, negative, mixed/can't tell). In all, 175,818 blog posts were judged by NIST assessors for this task. After TREC 2008, the general feeling was that the 150 topics represented a sufficient and high quality test collection for evaluating opinion-finding approaches. For this reason, it was decided that this task need not run in

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

[2] For instance, `http://blogpulse.com` and the former `http://summize.com`.

```
<top>
  <num> Number: 871

  <title> cindy sheehan

  <desc> Description:
  What has been the reaction to Cindy Sheehan and the
  demonstrations she has been involved in?

  <narr> Narrative:
  Any favorable or unfavorable opinions of Cindy Sheehan are
  relevant. Reactions to the anti-war demonstrations she has
  organized or participated in are also relevant.
</top>
```

Figure 1: Blog track 2006, opinion-finding task, topic 871.

Year	Topics
2006	851–900
2007	901–950
2008	1001–1050

Table 2: Topics for the opinion-finding tasks.

future TREC campaigns. Indeed, this test collection continues to be actively used for opinion-finding research outwith TREC, as can be seen from the large and growing list of publications described in the following section.

The effectiveness of retrieval systems on this task was assessed using standard IR evaluation metrics, such as Mean Average Precision (MAP), P@10, and r-Precision. However, for measuring the systems' ability to retrieve opinionated blog post documents, the threshold for relevance excluded documents that do not express an opinion.

3.2 Polarity Sub-Task

The judgements created by the NIST assessors in the opinion-finding task identified the polarity of the opinion expressed in each blog post. This allows investigations into the ability of systems to predict whether a document expresses a positive or a negative opinion about a blog post, as a natural refinement of the opinion-finding task. Indeed, both tasks share the same set of topics.

For TREC 2007, the polarity sub-task was formulated as a classification task, whereby for each retrieved documents, participants should predict the polarity of the document. For evaluation, R-Accuracy was used [30]. For TREC 2008, this task was reformulated as a ranking task—i.e., only blog posts which were relevant to the topic, and express a positive (resp. negative) opinion should be retrieved. In this case, the evaluation was naturally made using standard IR metrics, such as MAP.

3.3 Published Approaches

Opinion mining has a long history in the Natural Language Processing (NLP) community (for an excellent review of sentiment analysis, see Pang & Lee [41]). However, the opinion-finding task at TREC was the first to investigate the effectiveness of sentiment analysis as part of an IR ranking task. Over the three years of the opinion-finding task, it was noted that most participants approached this task as a re-ranking problem [30, 38, 40]. In the first stage, the systems aim to find as many relevant

blog posts as possible, regardless of their opinionated nature, while in the second stage, these blog posts are re-ranked using an opinion detection technique and an appropriate combination of scores.

To support a comparative evaluation of the opinion-finding approaches, in TREC 2008, we introduced the notion of *standard baselines* [40]. These five baselines are highly performing topical relevance systems (with no opinion-finding features enabled). Participants then deployed their two-stage opinion-finding approaches on top of the five standard baselines, allowing fair experimental comparisons without varying the underlying topical relevance retrieval system. For this reason, we view the standard baselines as an important feature of the opinion-finding test collection—indeed, they are actively being used in the literature for newly proposed opinion-finding approaches. In the following, we survey the effective techniques published in the literature for the opinion-finding task. These can be roughly organised into two main categories: classification-based and lexicon-based.

3.3.1 Classification-based Opinion-Finding

Classification-based opinion-finding approaches build a classifier using training data from sources known to contain either subjective (i.e., opinionated) or objective content. The trained classifier is then used to estimate the subjectiveness of blog posts in the target collection. This approach was investigated by Zhang et al. [60], who trained an SVM classifier using data from multiple sources. Subjective training data was obtained for each of the concepts identified from all 150 opinion-finding topics from two consumer review websites: RateItAll[3] and Epinions.com[4]. The trained classifier was then used to estimate the subjectiveness of individual sentences in blog posts from the Blogs06 collection. In order to estimate how on-topic the classified subjective sentences are—i.e., to what extent the expressed opinions are about the topic under consideration—a subsequent step was applied to look at the occurrence of query terms or concepts close to the identified subjective sentences. Their work was later refined based on a second-tier classification on top of the initially classified sentences. In particular, the output of the sentence-level SVM classifier was used to produce several document-level features for a decision tree classifier, which labels the entire documents [59].

Another effective classification-based approach was proposed by He et al. [14]. Their approach uses OpinionFinder [55], a subjectivity analysis system aimed to support NLP applications, to provide information about opinions expressed in text and also about who expressed them. OpinionFinder operates as a two-stage pipeline. The first stage performs general-purpose document processing (e.g., part-of-speech tagging, named entity identification, tokenisation, stemming, and sentence splitting), while the second stage is responsible for the subjectivity analysis itself. It employs a Naive Bayes classifier to distinguish between objective and subjective sentences. This classifier is trained on sentences automatically generated from a large corpus of unannotated data by two rule-based classifiers. He et al.'s approach computes an opinion score for each retrieved blog post based on the proportion of sentences in the blog post that OpinionFinder classifies as subjective, and also on the overall confidence of such classification [14].

Zhang et al. [61] proposed a simple classification-based opinion-finding technique. In particular, in their approach, documents are represented as vectors over opinionated words, drawn from a manually-annotated lexicon. An SVM classifier is then trained on a pool of documents induced from the TREC 2006 Blog track relevance assessments. Two variants were considered: topic-independent, with documents sampled from the relevance assessments of all 50 queries; and topic-dependent, with a different pool sampled from the assessments of each of the 50 queries. Their cross-validation results showed that the topic-dependent variant is significantly more accurate, suggesting that topically-related documents also share similar sentiment words.

3.3.2 Lexicon-based Opinion-Finding

Vechtomova [47] used a subjective lexicon manually derived from several linguistic resources. Using the relevance assessments from the opinion-finding tasks in TREC 2006 and 2007, each subjective term

[3] `http://www.rateitall.com`
[4] `http://www.epinions.com`

in the lexicon is scored based on the Kullback-Leibler (KL) divergence between its distribution in the set of opinionated blog posts in the collection and that in the set of all other blog posts (i.e., those judged as non-opinionated). In order to re-rank the blog posts retrieved by the topic-relevance baseline component, a modified implementation of BM25 is used: besides accounting for the frequency of the query terms in each blog post, it also considers the normalised KL scores of the subjective terms that co-occur with any query term within a window of 30 words in the blog post. By doing so, her approach addresses the requirement that relevant blog posts should contain expressed opinions towards the topic of interest.

Amati et al. [1] proposed an information-theoretic approach to automatically select the most effective terms from an opinionated lexicon. The opinionatedness of a given term in the lexicon is estimated based on the divergence of this term's distribution in a set of opinionated documents from that in a set of topically relevant yet non-opinionated documents. Terms more uniformly distributed among opinionated documents are preferred, as they are more likely to convey an opinion regardless of a particular topic. Finally, all the terms in the obtained vocabulary are submitted to a retrieval system as a query, so as to assign documents a topic-independent opinionated score.

Another effective approach was proposed by He et al. [13]. In their approach, however, the subjective lexicon is automatically derived from the target collection itself, hence alleviating the effort required to build it manually. Firstly, from the list of all terms in the collection ranked by their within-collection frequency (i.e., the number of blog posts in which the term occurs in the collection) in descending order, a skewed query model is applied to filter out those that are too frequent or too rare [6]. This aims to remove terms with too little or too specific information and which cannot be interpreted as generalised, query-independent opinion indicators. Using a training set of queries, the remaining terms from the list are weighted based on the divergence of their distribution in the set of opinionated documents retrieved for these queries against that in the set of relevant documents retrieved for the same set of queries. The top weighted opinionated terms in the lexicon are then submitted as a query. The score produced for this opinionated query is then used as the opinion score of the blog post, and is combined with its relevance score (obtained from a topic-relevance baseline) in order to generate the final ranking. This approach was also shown to be suitable for estimating the polarity of blog posts (see Section 3.2).

An adaptation of this approach was proposed by Santos et al. [42]. In their approach, the automatically built subjective lexicon is used to estimate the subjectiveness of individual sentences in the retrieved blog posts. This provides a coarser grained estimation of the opinionated content in these blog posts—as opposed to considering subjective terms outside their surrounding context—while relying on a light-weight alternative to using a classifier, such as OpinionFinder. After identifying subjective sentences, the retrieved blog posts are then scored based on the occurrence of the query terms in proximity to these sentences. By doing so, they focus on opinionated content targeting the topic of interest instead of any given expression of opinion. The produced opinion scores are then combined with the corresponding relevance scores and the retrieved blog posts are re-ranked.

Seki and Uehara [44] devised a language modelling opinion-finding approach based on the notion of subjective triggers. The basic idea is that the target concept in the query can *trigger* certain subjective expressions. Trigger patterns were automatically identified from Amazon review data, and the resulting trigger model was interpolated to a baseline n-gram language model. Their experiments showed that substantial improvements can be attained on top of this baseline.

Another effective lexicon-based approach was proposed by Lee et al. [37]. In their approach, a lexicon is built based on a generative model of words from two distinct resources: the Amazon's product review corpus, serving as a subjective resource, and the Amazon's product specification corpus, serving as an objective resource. Words in the lexicon are weighted based on the combination of the two models, with the estimation of subjectivity of a single word according to SentiWordNet[5] used as its prior probability of being opinionated. After building the lexicon, blog posts retrieved by a topic-relevance baseline are scored and re-ranked based on the length-normalised sum of the opinion scores of individual words in each blog post. Besides being effective in finding general opinions, this approach was also shown to be effective for polarity detection, by employing the actual ratings in the Amazon's product review corpus as an indicator of the semantic orientation of the expressed opinions.

[5] `http://sentiwordnet.isti.cnr.it`

Differently from the aforementioned two-stage opinion-finding approaches, Zhang and Ye [58] proposed to unify the estimations of topical relevance and opinionatedness into a single-stage generative model for opinion-finding. In their approach, the probability of a document being opinionated is conditioned on the observation of the query terms. As a baseline in their investigation, they employed a typical two-stage approach, in which a relevance score and an opinionated score for each blog post are linearly combined. Through a comprehensive analysis, they showed that the proposed unified model outperforms the independent estimations of topical relevance and opinionatedness, with significant improvements over the best TREC systems.

In a similar vein, Huang and Croft [16] proposed an alternative single-stage opinion-finding approach. However, instead of modelling the document generation process, they based their approach upon relevance models [23]. In particular, a query-independent 'sentiment' expansion technique was proposed for expanding the initial query with opinionated words from multiple sources, including seed words (e.g., 'good', 'nice', 'bad', 'poor') and review data. The weight of each expansion term was learnt based on the contribution of this term to the opinion retrieval performance of a set of training queries. Alternatively, they proposed a query-dependent sentiment expansion based on (pseudo-)relevance feedback, in order to select opinion words that co-occur with the original query terms in a (pseudo-)feedback set. The combination of their query-independent and query-dependent sentiment expansion techniques was shown to outperform both mechanisms individually.

Finally, the lexicon-based approach of Gerani et al. [11] uses the proximity of opinionated terms and query terms at a term-level (this contrasts with the sentence-level approach of Santos et al. [42]). This is achieved by a kernel to calculate the "opinion density" at each point in a document. Three kernels (Gaussian, Laplace and Triangular) are investigated, with the Laplace kernel being found most effective, showing improvements over all of the standard baselines introduced in the TREC 2008 opinion-finding task.

3.4 Summary

With the breadth of published approaches for opinion-finding, as well as a participation of no less than 45 different groups over the three years that it ran at the Blog track, we are pleased with the success of this task, and the new research opportunities it has generated. Indeed, the opinion-finding task lies at the intersection of NLP, machine learning and IR, and this has been mirrored in the wide backgrounds of groups working on the task.

While the opinion-finding task has now ended in the Blog track, the reusable test collection formed has been and is still being widely used in many publications. Moreover, related tasks in question answering and summarisation with the Blogs06 corpus have been spawned as part of the Text Analysis Conference (a TREC sibling, also under the auspices of NIST).

4 Blog Distillation

While blogs may be written in either a personal or a more official/professional role, bloggers often have topics which interest them, and about which they blog regularly. At the same time, users often wish to find blogs that regularly blog about topics that interest them (For instance, the authors read primarily information retrieval blogs). Indeed, in a study of the query log of a blog search engine, Mishne & de Rijke [36] identified *concept queries* where users seemed to be looking for blogs to subscribe to these blogs using their RSS reader software.

A primary way of finding new interesting blogs comes through the blogroll lists present in most blogs. Bloggers often add to the right hand side of their blog a list of other related blogs that they read or endorse. Readers of their blog may then use this list to find other interesting blogs. There also exist a number of blog directories, containing submissions of blogs, which users can use to find blogs of interest.

Initially, blog search engines only provided blog post search, i.e. a list of relevant blog posts returned in response to a user query. Moreover, many manually-categorised blog directories exist, such as Blogflux and Topblogarea to name but a few. This is reminiscent of the prevalence of the

Year	Topics	Corpus	Notes
2007	951–995	Blogs06	
2008	1051–1100	Blogs06	Graded
2009	1101-1150	Blogs08	Faceted

Table 3: Topics for the Blog Distillation tasks.

early Web directories (c.f. Yahoo!) before Web search matured, and suggests that there is indeed an underlying user task that needs to be researched [17]. However, both Google[6] and Technorati[7] now provide blog search services, where blogs related to the query are retrieved.

4.1 Task Definition

The aim of the *blog distillation task*, introduced in TREC 2007, was for systems to suggest relevant blogs in response to a query. The task can be summarised as *Find me a blog with a principle, recurring interest in X*. For a given target X, systems should suggest blogs that are principally devoted to X over the timespan of the blog, and would be recommended to subscribe to as an interesting blog about X (i.e. a user may be interested in adding it to their RSS reader). An example topic is shown in Figure 2. The blog distillation task contrasts from the opinion-finding task, in that blogs rather than blog posts are retrieved. In particular, both the Blogs06 and Blogs08 collections provide mappings from blog posts to blogs.

```
<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 2: Blog track 2007, blog distillation task, topic 994.

A total of 150 blog distillation topics have been created since TREC 2007. Table 3 details the relevant topic numbers. Of note is that TREC 2007 and TREC 2008 used the Blogs06 corpus, while TREC 2009 used the Blogs08 corpus. In the TREC 2007 task, binary relevance gradings were used. In TREC 2008, 3 levels of relevance were used (not relevant, relevant, highly relevant). In TREC 2009, a new refined version of this task was introduced, as described in the next section.

[6] http://blogsearch.google.com
[7] http://technorati.com

4.2 Faceted Task Definition

In its original form, the blog distillation task did not address the "quality" aspect of the retrieved blogs. However, a position paper by Hearst et al. [15] described how a blog search engine might present blogs using an exploratory search interface. In such an interface, the user may be presented with facets that allow the filtering of blogs according to various attributes. Such facet attributes might represent the opinionated nature of the blog, the trustworthiness of its authors, or its style of writing or genre.

For TREC 2009, we introduced a refinement of the blog distillation task to investigate retrieval approaches facilitating such exploratory search. Indeed, the *faceted blog distillation task* is the first operationalisation in a TREC-setting of an exploratory search scenario. In this task, the aim is to retrieve a ranking of blogs having a recurring and principal interest in a given topic X, but only those blogs that satisfy the active facet inclination(s). For example, a user might be interested in blogs to read about a topic X, but where the blogger expresses opinionated viewpoints, backed up by an in-depth analysis. Hence, this task can therefore be summarised as "Find me a *suitable* blog with a principal, recurring interest in X", where the appropriate nature of the blogs is characterised through the active facet inclination(s).

For TREC 2009, we defined an initial set of three facets of varying difficulty, which were all assumed to have binary inclinations (where the inclination describes what blogs are required):

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

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

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

It is of note that many other exploratory search scenarios use structured data (e.g. online shopping) where the facet attributes are explicit (e.g. price). In contrast, the facets proposed above are latent attributes, rather than explicitly determined.

One appropriate facet was chosen for each of the topics developed for TREC 2009 (see Table 3). In particular, the facet Opinionated was chosen for 21 topics, the facet Personal was chosen for 10 topics, and the facet In-depth was chosen for 19 topics. Finally, during judging, blogs were judged as to their relevance, and whether they matched the required facet inclination for the topic.

In the following section, we detail the published approaches for blog distillation from the literature. Due to its relative recency, for approaches for the faceted blog distillation task, we refer the reader to the TREC 2009 Proceedings [52].

4.3 Published Approaches

As mentioned above, among all tasks investigated thus far within the TREC Blog track, the blog distillation task has been the only one to address the problem of searching for entire blogs instead of individual blog posts. This granularity issue is also what essentially differentiates the most effective approaches for blog distillation reported in the first three editions of this task. While some approaches address this task as a traditional document search task but with "large" documents (i.e., blogs), others cast it as a problem of searching for collections of "small" documents (i.e., blog posts) and then inferring a ranking of blogs. Moreover, two separate themes emerge in the literature, which we detail in two distinct sections below.

4.3.1 Blog Distillation as a Resource Selection Problem

The view of blogs as collections of posts directly links the blog distillation task to the resource selection problem in distributed information retrieval [7], in which the goal is to select the collections (or resources) more likely to contain documents relevant to a given query. Elsas et al. [10] were among the first to explore the link between the two tasks by proposing different models for representing blogs, to make an explicit distinction between using the large documents (blogs) vs. the small documents (blog posts) views of blog search. The first representation, called the Large Document (LD) model, treats a blog as a concatenation of all its posts. An additional enhancement considers not only the HTML content of posts, but also their different fields (e.g. blog title, post title) from the blog's syndicated XML feeds [2]. This combined representation is then used to rank blogs by their posterior probability given the query, with the query likelihood of each component estimated using a full dependence model in order to account for term dependencies [35]. In the second representation, known as the Small Document (SD) model, blogs are seen as a collection of blog posts, as in a typical resource selection problem. Accordingly, they adapt a state-of-the-art resource selection algorithm, which attempts to estimate the number of relevant documents in a remote collection by sampling this collection. Analogously, in the SD model, a blog is scored based on a combination of the individual scores of its posts. Again, individual posts are represented using different fields from both their HTML and XML representations. Besides the field-based query likelihood computed for each post, a query-biased centrality component is used to infer how well the post represents the language model of the whole blog with respect to the query terms [10]. The comparison between the two models shows that LD performs markedly better than SD, except for when the centrality component is integrated, in which case SD was shown to perform significantly better [10].

Besides different retrieval models, Elsas et al. [10] also proposed a query expansion mechanism based on an external resource. In their approach, Wikipedia articles are ranked for a given query, and two overlapping portions of the resulting ranking are considered: a *working* portion, comprising the top w retrieved articles, and a *relevant* portion, comprising the top r articles, with $r < w$. The anchor phrases—the clickable text in a hyperlink—in the working portion of the ranking and that refer to articles in the relevant portion are scored based on two criteria: (1) their frequency in the working articles, and (2) the rank $(1 \dots r)$ of their referred articles in the relevant portion. From these, the top 20 anchor phrases are selected to expand the original query. The results showed that this simple mechanism can markedly improve baseline rankings using both LD and SD representations.

Seo & Croft [45] also approached blog distillation as a resource selection problem, by seeing blogs as a collections of blog posts. Similarly to the approach of Elsas et al., they also consider different representations of blogs, namely, a Global Representation (GR), and an alternative representation, called Pseudo-Cluster Selection (PCS). The GR model simply treats blogs as a concatenation of all their posts, as in the LD model. PCS is analogous to the SD model of Elsas et al., but is based on a different principle. In PCS, a blog is seen as a query-dependent cluster containing only highly ranked blog posts for a given query. Considered separately, GR was shown to outperform PCS on the 2007 topic set of the blog distillation task. Additionally, a strategy that combines the two models was shown to outperform both individually, hence demonstrating their complementary characteristics [45]. Indeed, the global nature of GR helps uncover the prevailing topics covered by a blog instead of a multitude of relatively less important topics, whereas the selection strategy employed by PCS mitigates the potential problem of a blog being overly represented by a few, long blog posts. Finally, to avoid the issue of operating with distinct indices (GR uses an index of feeds, while PCS is based on an index of posts), Seo and Croft proposed an alternative to GR. To play the role of penalising topical diversity while reducing the overhead of having a second index structure, they create a query-independent version of PCS, by randomly sampling blog posts from a blog. Furthermore, in order to focus on the temporal aspect of blogs, where those with more recent blog posts on a given topic are more likely to be relevant to the topic, this sampling is biased towards recently added blog posts. This alternative version is shown to perform comparably to the GR model [45].

A third approach to blog distillation built upon the two previously described approaches. In particular, Lee et al. [25] adapted the SD model of Elsas et al. [10]—renamed the Global Evidence Model (GEM)—and the PCS model of Seo & Croft [45]—renamed the Local Evidence Model (LEM)—

in the context of the risk minimisation framework of Lafferty & Zhai [22]. In practice, both GEM and LEM are implemented identically, except that GEM considers every post in a given blog, whereas LEM only considers the top retrieved blog posts for a given query. Besides being based on a different framework, their approach directly addresses two weaknesses of both SD and PCS when considered individually. Firstly, to overcome the problem of blogs being overly represented by a few long posts, all blog posts are considered equally important to the blog they belong to, which is expressed in their probabilistic framework as a uniform probability of posts being retrieved given their blog. Secondly, to avoid a bias towards prolific blogs (i.e., those with a distinctively large number of blog posts), the score of a given blog is computed as the average score of its posts. Finally, the combination of the proposed models was further improved by a diversity-oriented query expansion technique. Differently from a traditional application of pseudo-relevance feedback techniques to this task—which would probably consider the top retrieved posts for a given query in order to select expansion terms—their approach considers the top retrieved posts from the top retrieved blogs as the pseudo-relevance feedback set. Since a blog covers a broader range of topics—or even different perspectives of a single topic—when compared to a single blog post, their approach provides a richer vocabulary around the topic of the query, which has the potential to produce a more effective expanded query.

4.3.2 Blog Distillation as an Expert Search Problem

A different class of approaches to blog distillation explores the similarities of this task to the expert search task [3]. In an expert search task, the goal is to find people with relevant expertise on a particular topic of interest. Analogously, the estimated relevance of blog posts to a given query can be seen as an indication of the interest of the bloggers who authored these posts with respect to the topic of the query. Macdonald and Ounis [28] were the first to propose tackling blog distillation as an expert search problem, by adapting their expert search model—the so-called Voting Model [27]—to the task of searching for "experts" in the blogosphere (i.e., bloggers or, in this case, their blogs). The Voting Model is based on the notion of profiles. The profile of a blogger contains all blog posts authored by this blogger. The blog posts in the profiles of all bloggers in the test collection can be used to rank these bloggers (i.e., their blogs) in response to a query according to their "expertise" to the topic of the query. The basic idea is that blog posts retrieved for a given topic that belong to the profile of a blogger are considered as votes for the relevance of that blogger to the topic. The Voting Model defines many voting techniques, which provide different ways of converting a ranking of blog posts into a ranking of blogs, many of which were shown to be effective for blog distillation [28]. Additionally, techniques intended to enhance the underlying ranking of blog posts were shown to yield an improved blog distillation effectiveness using the Voting Model. For instance, taking into account the proximity of query terms in the retrieved posts was shown to be beneficial, as well as expanding the original query using Wikipedia [28]. Besides enhancing the ranking of blog posts, techniques applied to the ranking of blogs brought additional improvements. These include a technique to favour blogs with a recurrent interest in the topic of the query, which estimates how well the posts in a single blog are spread throughout the time frame covered by the test collection. Also, a technique to counterbalance any bias towards prolific bloggers (i.e., those with large profiles and, hence, more likely to be retrieved for any query) was also shown to further improve the retrieval effectiveness [29].

This connection to the expert search task was also explored by Balog et al. [5] using the language modelling framework. In their approach, two expert search models [4] were adapted to the blog distillation task. In the Blogger Model (BM), the probability of a query being generated by a given blog is estimated by representing this blog as a multinomial distribution over terms. In the Posting Model (PM), this probability is estimated by combining the estimated probabilities of the individual blog posts in this blog generating the query. In both models, the prior probability of choosing a given blog is considered uniform, as well as the probability of choosing a post given its blog. Their results showed that BM significantly outperforms PM in this task, contrarily to the relative performance observed for their counterpart models in the expert search task [4]. This reinforces the observation that, differently from the expert search task, the blog distillation task requires bloggers to not only write about the topic of interest, but to do so on a focused and recurrent basis.

Keikha & Crestani [19] applied an ordered weighted average operator (OWA) for aggregating blog post scores. OWA provides a parametrised class of mean aggregation operators, that can generate an OR (Max), AND (Min) and other operators. In an experimental comparison with both voting techniques and language models, they showed that performance for blog search could be further enhanced.

Moreover, both Keikha et al. [18] and Weerkamp et al. [54] propose extensions to the language modelling-based approaches for blog distillation. In particular, Keikha et al. [18] uses a tri-partite graph of posts, blogs and terms, encapsulating term occurrences, associations between blogs and blog posts, as well as hyperlinks between the posts. They then perform a random walk on this graph to identify the most important blogs for a given query. Performance improvements are observed over the language modelling approach of Balog et al. [5]. Lastly, Weerkamp et al. [54] extend the language modelling approach with priors for temporal and social structure of blogs.

4.4 Summary

The blog distillation task tackles an important information need of blog searchers in the blogosphere, namely to search for blogs to read on a regular basis. With the refinements introduced by the new faceted version of the task, it also represents the first operationalisation of an exploratory search task within the TREC evaluation paradigm.

Many of the approaches for this task are unusual in that the retrieval units are often not what the IR system is initially ranking (e.g. small document-based models). The breadth of published models attests the research potential of the task, and with several PhD students currently working on this topic, we look forward to seeing more advanced models and features being published in the future for the faceted blog distillation task.

5 Top News

A poll by Technorati found that 30% of bloggers considered that they were blogging about news-related topics [34]. Similarly, Mishne & de Rijke [36] showed a strong link between blog searches and recent news - indeed almost 20% of searches for blogs were news-related. As an illustration, Thelwall [46] explored how bloggers reacted to the London bombings, showing that bloggers respond quickly to news as it happens. Furthermore, both König et al. [20] and Sayyadi et al. [43] have exploited the blogosphere for event analysis and detection, showing that news events can be detected within the blogosphere.

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

For these two scenarios, we developed the top news identification task of the TREC 2009 Blog track. This task had two aims: firstly, to evaluate the ability of systems to automatically identify the top news stories on a given day, as an editor would do, but using only evidence from the blogosphere; secondly, to provide related blog posts covering diverse perspectives of that news story[8]. In the following, we define the top news stories identification task, and describe the main approaches to the task.

5.1 Task Definition

In TREC 2009, we devised the top news stories identification task, with the purpose of exploring how accurately top news could be determined using the blogosphere [32]. In particular, the aim of the

[8]Search result diversification is currently a popular topic as exemplified by the diversity task of the TREC 2009 Web track [8].

task was for systems to use the Blogs08 corpus to suggest a ranking of news articles where the most important, "news-worthy" articles were ranked first. During evaluation, these could then be compared to which news articles were deemed to be editorially important on each given day.

To operationalise this task, we picked 55 dates in the timespan of Blogs08, where interesting events had occurred. Furthermore, a set of 102,000 headlines were obtained from the New York Times. Participant systems were asked to rank 100 headlines for each date, by their predicted importance. As a second angle of the task, for each ranked news headline, 10 blog posts were also ranked, to provide a diverse set of related blog posts. In doing so, the aim was for systems to give users an overview of the main opinions, forms or aspects of blog posts concerning that news article.

Evaluation was performed in two stages. Firstly, by pooling the important headlines suggested by systems for each date, assessors identified the most important headlines from an editorial perspective. This allowed the evaluation of systems for identifying top new stories, using traditional IR evaluation measures such as MAP, P@10, etc. In the second evaluation stage, the related blog posts suggested by systems were pooled for a subset of the relevant headlines. These were then assessed for topical relevance, and also grouped into "aspects". For instance, blog posts about the 2008 Academy Awards were grouped into aspects such as 'live blogs', 'prediction', 'aftermath', etc. To evaluate this aspect of system performance, we used the α-nDCG diversity measure [9].

5.2 Approaches

With the recency of this task, only a few approaches have been published outwith the TREC 2009 proceedings. We review those published thus far below. For other approaches, we refer the reader to the TREC 2009 proceedings [52].

In particular, McCreadie et al. [33] devise a voting approach for identifying relevant headlines based on the Voting Model [29]. In particular, blog posts are ranked for each headline. Then based on the number of retrieved posts for each headline, the importance of each headline on that day is inferred. Additionally, by examining the historical importance of a headline over time found that significant improvements in effectiveness could be achieved.

Moreover, Lee at al. [24] proposed a language modelling approach to rank news stories by their importance on a day of interest using evidence from the blogosphere. In particular, they use clustering to create multiple topic models for a day, and compare these to a headline model generated from the top retrieved blog posts for that headline. Finally, additional temporal evidence is used in the form of a headline prior.

5.3 Summary

The first year of the top news stories identification task raised interesting research and perspectives on the task. Moreover, the TREC 2009 task was a pilot task, where the task was operationalised in a retrospective setting (i.e. the Blogs08 corpus was treated as a whole, rather than as a time stream). In future iterations of this task, it will be modelled more in the style of an online event detection [57], whereby the corpus is treated as a stream, and evidence after the query date cannot be used.

6 Conclusions

In this article, we have motivated and summarised the work conducted during the first four years of the TREC Blog track. In the following, we provide further information, and several concluding remarks and pointers in relation to the Blog track, its current status and its future.

6.1 Summary of TREC Blog track 2006-2009

Over the past four years, the TREC Blog track has investigated three information retrieval tasks within the context of the blogosphere, namely opinion-finding, blog distillation and top news identification. A total of 59 different groups have participated in the Blog track since its inception. Techniques for

social search in general, and blog search in particular, are increasingly being proposed and published in the literature. The Blog track has played an important role in initiating research, creating resources and facilitating the formation of a community of researchers for tackling such multi-disciplinary search tasks. Indeed, as coordinators, we are pleased with the volume of published research emanating from the test collections of the Blog track, the influence of the track in the opening up of information retrieval research to new challenges and topics, and the fostering of a sustainable IR community in blog search.

6.2 Further Reading

In this article, we have identified work published in refereed publications describing successful and effective approaches for the tasks tackled within the Blog track framework. However, there are also several other approaches described by participants in the TREC proceedings [49, 50, 51, 52][9]. In particular, the TREC Blog track overview papers provide a detailed description of the approaches deployed by the participating groups in the past four years [30, 32, 38, 40].

One of the characterising features of the blogosphere is the abundance and severity of spam. In [39], we analysed the difficulty of both opinion-finding and blog distillation topics, in terms of median retrieval performance, and the amount of spam blog posts retrieved. The correlation between the system performance and tendency to retrieve spam was further analysed in [31]. As described in Section 2, the Blogs06 collection had been interjected with a number of assumed spam blogs (splogs). To encourage researchers to investigate the effect of spam on blog search, as well as to develop more effective spam detection techniques for the blogosphere, we have recently released the Blogs06 labelled spam dataset[10].

6.3 Future Directions

The Blog track will continue in TREC 2010. In particular, the faceted blog distillation and top news stories identification tasks will continue in similar forms. For the faceted blog distillation task, more topics will be developed. Given the limited TREC resources, we intend to examine the usefulness of crowdsourcing (such as Amazon Mechanical Turk) for the development and assessment of a large number of additional topics. By using crowdsourcing, we aim to establish best practises and guidelines for collecting relevance judgements through this paradigm, and assessing their quality in the confines of the Blog track setting. For the top news stories identification task, we will refine the task, so that it becomes of an online event detection type, instead of its current retrospective nature. Given the additional assessment requirements of the top news stories identification task, we also intend to use crowdsourcing to complement the usual TREC assessment procedure.

A common feature of the Web 2.0 is the micro-blogging status updates, such as that used by Facebook and Twitter. We hope in the future that datasets will become available to facilitate the creation of sharable, reusable test collections within this area. Unfortunately, currently, restrictions by the micro-blogging services prevent research formulating under the TREC umbrella, as datasets cannot be distributed. We are actively lobbying for the easing of such restrictions.

Acknowledgements

Firstly, we would like to extend our thanks to NIST for supporting the Blog track for the last four years. We are also thankful to Gilad Mishne and Maarten de Rijke for joining us in organising the TREC 2006 Blog track. Finally, we are extremely grateful to the 59 participating groups of the TREC Blog track over the past four years, including those who devoted their valuable time making relevance assessments.

[9] Available online from `http://trec.nist.gov/proceedings/proceedings.html`

[10] `http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG`

Disclaimer

The mention of companies or products in this paper should in no way be construed as indicating that such products or companies are endorsed by NIST or are recommended by NIST or that they are necessarily the best companies or products for the purposes described.

References

[1] G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi. Automatic construction of an opinion-term vocabulary for ad hoc retrieval. In *Proceedings of the 30th European Conference on IR Research on Advances in Information Retrieval (ECIR 2008)*, pages 89–100, 2008.

[2] J. Arguello, J. Elsas, J. Callan, and J. Carbonell. Document representation and query expansion models for blog recommendation. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*. AAAI, 2008.

[3] P. Bailey, N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC-2007 Enterprise track. In *Proceedings of the 16th Text REtrieval Conference (TREC 2007)*, 2007.

[4] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, pages 43–50. ACM, 2006.

[5] K. Balog, M. de Rijke, and W. Weerkamp. Bloggers as experts: Feed distillation using expert retrieval models. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, pages 753–754. ACM, 2008.

[6] F. Cacheda, V. Plachouras, and I. Ounis. A case study of distributed information retrieval architectures to index one terabyte of text. *Information Processing and Management*, 41(5):1141–1161, 2005.

[7] J. Callan. Distributed information retrieval. In W. B. Croft, editor, *Advances in Information Retrieval*, chapter 5, pages 127–150. Kluwer Academic Publishers, 2000.

[8] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web track. In *Proceedings of the 18th Text REtrieval Conference (TREC 2009)*, 2010.

[9] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2008)*, pages 659–666. ACM, 2008.

[10] J. L. Elsas, J. Arguello, J. Callan, and J. G. Carbonell. Retrieval and feedback models for blog feed search. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, pages 347–354. ACM, 2008.

[11] S. Gerani, M. Carman, and F. Crestani. Proximity based opinion retrieval. In *Proceedings of the 33rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2010)*, 2010.

[12] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web (WWW 2004)*, pages 491–501. ACM, 2004.

[13] B. He, C. Macdonald, J. He, and I. Ounis. An effective statistical approach to blog post opinion retrieval. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008)*, pages 1063–1072. ACM, 2008.

[14] B. He, C. Macdonald, and I. Ounis. Ranking opinionated blog posts using OpinionFinder. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, pages 727–728. ACM, 2008.

[15] M. A. Hearst, M. Hurst, and S. T. Dumais. What should blog search look like? In *Proceedings of the 1st International Workshop on Search in Social Media (SSM 2008)*, pages 95–98. ACM, 2008.

[16] X. Huang and W. B. Croft. A unified relevance model for opinion retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM 2009)*, pages 947–956. ACM, 2009.

[17] A. Java, P. Kolari, T. Finin, A. Joshi, and T. Oates. Feeds That Matter: A Study of Bloglines Subscriptions. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*. Computer Science and Electrical Engineering, University of Maryland, Baltimore County, March 2007.

[18] M. Keikha, M. J. Carman, and F. Crestani. Blog distillation using random walks. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2009)*, pages 638–639. ACM, 2009.

[19] M. Keikha and F. Crestani. Effectiveness of aggregation methods in blog distillation. In *Proceedings of the 8th International Conference on Flexible Query Answering Systems (FQAS 2009)*, pages 157–167. Springer-Verlag, 2009.

[20] A. C. König, M. Gamon, and Q. Wu. Click-through prediction for news queries. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2009)*, pages 347–354. ACM, 2009.

[21] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proceedings of the 12th international conference on World Wide Web (WWW 2003)*, pages 568–576. ACM, 2003.

[22] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pages 111–119. ACM, 2001.

[23] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2001)*, pages 120–127. ACM, 2001.

[24] Y. Lee, H.-Y. Jung, W. Song, and J.-H. Lee. Mining the blogosphere for top news stories identification. In *Proceedings of the 33rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2010)*, 2010.

[25] Y. Lee, S.-H. Na, and J.-H. Lee. An improved feedback approach using relevant local posts for blog feed retrieval. In *Proceeding of the 18th ACM conference on Information and knowledge management (CIKM 2009)*, pages 1971–1974. ACM, 2009.

[26] C. Macdonald and I. Ounis. The TREC Blogs06 collection: creating and analysing a blog test collection. Technical Report TR-2006-224, Department of Computing Science, University of Glasgow, 2006.

[27] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM 2006)*, pages 387–396. ACM, 2006.

[28] C. Macdonald and I. Ounis. Key blog distillation: ranking aggregates. In *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM 2008)*, pages 1043–1052. ACM, 2008.

[29] C. Macdonald and I. Ounis. Searching for expertise: Experiments with the voting model. *Computer Journal: Special Focus on Profiling Expertise and Behaviour*, 52(7):729–748, 2009.

[30] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2007 Blog track. In *Proceedings of the 16th Text REtrieval Conference (TREC 2007)*, 2007.

[31] C. Macdonald, I. Ounis, and I. Soboroff. Is spam an issue for opinionated blog post search? In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2009)*, pages 710–711. ACM, 2009.

[32] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2009 Blog track. In *Proceedings of the 18th Text REtrieval Conference (TREC 2009)*, 2009.

[33] R. M. C. McCreadie, C. Macdonald, and I. Ounis. News article ranking: Leveraging the wisdom of bloggers. In *Proceedings of the 9th International Conference on Computer-Assisted Information Retrieval (RIAO 2010)*, 2010.

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

[35] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pages 472–479. ACM, 2005.

[36] G. Mishne and M. de Rijke. A study of blog search. In *Proceedings of the 28th European Conference on Information Retrieval (ECIR 2006)*, pages 289–301. Springer, 2006.

[37] S.-H. Na, Y. Lee, S.-H. Nam, and J.-H. Lee. Improving opinion retrieval based on query-specific sentiment lexicon. In *Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval (ECIR 2009)*, pages 734–738. Springer-Verlag, 2009.

[38] I. Ounis, C. Macdonald, M. de Rijke, G. Mishne, and I. Soboroff. Overview of the TREC 2006 Blog track. In *Proceedings of the 15th Text REtrieval Conference*, 2006.

[39] I. Ounis, C. Macdonald, and I. Soboroff. On TREC Blog track. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2008)*. AAAI, 2008.

[40] I. Ounis, C. Macdonald, and I. Soboroff. Overview of the TREC 2008 Blog track. In *Proceedings of the 17th Text REtrieval Conference (TREC 2008)*, 2008.

[41] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[42] R. L. T. Santos, B. He, C. Macdonald, and I. Ounis. Integrating proximity to subjective sentences for blog opinion retrieval. In *Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval (ECIR 2009)*, pages 325–336. Springer, 2009.

[43] H. Sayyadi, M. Hurst, and A. Maykov. Event detection and tracking in social streams. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2009)*. AAAI, 2009.

[44] K. Seki and K. Uehara. Adaptive subjective triggers for opinionated document retrieval. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM 2009)*, pages 25–33. ACM, 2009.

[45] J. Seo and W. B. Croft. Blog site search using resource selection. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008)*, pages 1053–1062. ACM, 2008.

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

[47] O. Vechtomova. Facet-based opinion retrieval from blogs. *Information Processing and Management*, 46(1):71–88, 2010.

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

[49] E. M. Voorhees and L. P. Buckland, editors. *Proceedings of the 15th Text REtrieval Conference (TREC 2006)*, 2007.

[50] E. M. Voorhees and L. P. Buckland, editors. *Proceedings of the 16th Text REtrieval Conference (TREC 2007)*, 2008.

[51] E. M. Voorhees and L. P. Buckland, editors. *Proceedings of the 17th Text REtrieval Conference (TREC 2008)*, 2009.

[52] E. M. Voorhees and L. P. Buckland, editors. *Proceedings of the 18th Text REtrieval Conference (TREC 2009)*, 2010.

[53] E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[54] W. Weerkamp, K. Balog, and M. de Rijke. Finding key bloggers, one post at a time. In *Proceedings of the 18th Conference on Artificial Intelligence (ECAI 2008)*, pages 318–322. IOS Press, 2008.

[55] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35. Association for Computational Linguistics, 2005.

[56] P. Winn. State of the Blogosphere, introduction, 2008. `http://technorati.com/blogging/article/state-of-the-blogosphere-introduction`.

[57] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 1998)*, pages 28–36. ACM, 1998.

[58] M. Zhang and X. Ye. A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2008)*, pages 411–418. ACM, 2008.

[59] W. Zhang, L. Jia, C. Yu, and W. Meng. Improve the effectiveness of the opinion retrieval and opinion polarity classification. In *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM 2008)*, pages 1415–1416. ACM, 2008.

[60] W. Zhang, C. Yu, and W. Meng. Opinion retrieval from blogs. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM 2007)*, pages 831–840. ACM, 2007.

[61] X. Zhang, Z. Zhou, and M. Wu. Positive, negative, or mixed? Mining blogs for opinions. In *Proceedings of the 14th Australasian Document Computing Symposium (ADCS 2009)*, 2009.