Foundations and Trends® in
Information Retrieval
Vol. 6, No. 1 (2012) 1–125

DOI: 10.1561/1500000026

now
the essence of knowledge

Information Retrieval on the Blogosphere

By Rodrygo L. T. Santos, Craig Macdonald,
Richard McCreadie, Iadh Ounis
and Ian Soboroff

Contents

Foundations and Trends® in
Information Retrieval
Vol. 6, No. 1 (2012) 1–125

DOI: 10.1561/1500000026

now
the essence of knowledge

Information Retrieval on the Blogosphere

Rodrygo L. T. Santos, Craig Macdonald, Richard McCreadie, Iadh Ounis[1] and Ian Soboroff[2]

[1] *University of Glasgow, UK,*
{rodrygo,craigm,richardm,ounis}@dcs.gla.ac.uk
[2] *National Institute of Standards and Technology, USA,*
ian.soboroff@nist.gov

Abstract

Blogs have recently emerged as a new open, rapidly evolving and reactive publishing medium on the Web. Rather than managed by a central entity, the content on the blogosphere — the collection of all blogs on the Web — is produced by millions of independent bloggers, who can write about virtually anything. This open publishing paradigm has led to a growing mass of user-generated content on the Web, which can vary tremendously both in format and quality when looked at in isolation, but which can also reveal interesting patterns when observed in aggregation. One field particularly interested in studying how information is produced, consumed, and searched in the blogosphere is information retrieval. In this survey, we review the published literature on searching the blogosphere. In particular, we describe the phenomenon of blogging and the motivations for searching for information on blogs. We cover

both the search tasks underlying blog searchers' information needs and the most successful approaches to these tasks. These include blog post and full blog search tasks, as well as blog-aided search tasks, such as trend and market analysis. Finally, we also describe the publicly available resources that support research on searching the blogosphere.

1

Introduction

The rise of the blogosphere has brought much attention in recent years toward this unique subset of the World Wide Web. In this section, we discuss the publishing phenomenon that has driven the growth of the blogosphere, with an emphasis on what makes it such an interesting experimental testbed for researchers in several fields including natural language processing, machine learning, and information retrieval.

1.1 Social Media

The last decade has witnessed a tremendous shift in publishing power. In particular, the Web has influenced not only the way information is distributed and consumed but, essentially, the way it is produced. Mainstream publishers now face a surge in user-generated content — in an unprecedented scenario, virtually every individual with an Internet connection becomes a potential information provider. Arguably, the act of blogging has played a major role in this paradigm shift [187], leading to not just the rise of grassroots journalism [67], but the provision of channels for anyone to espouse opinions [184], even if it does not guarantee an audience [47].

Although online communities have been around since the early days of the Internet — mainly in the form of newsgroups and discussion boards — it was only in the late 1990s that blogging began gaining in popularity as a means of self-expression, particularly with the advent of tools that facilitate the publishing process, as well as the inception of major blog hosting services [24, 25], such as Blogger[1] and Wordpress.[2] These enabled a much larger group of individuals to start blogging about practically anything and to interact with others sharing similar interests but possibly rather different points of view. This publishing phenomenon led to the formation of an increasingly growing network of self-publishers and their readership, with one of the major blog search engines currently tracking over 182 million blogs.[3] Of course, the blogosphere does not represent the entirety of online networked communities [47], with more social sites such as MySpace, Facebook, Google+, and Twitter all being heavily inspired by the blogosphere.

1.2 What is a Blog?

A *blog* (short for weblog) is a Web site generally authored by a single individual — known as a *blogger* — and updated on a regular basis. In terms of content organization, a typical blog comprises three main components [24, 25], depicted in Figure 1.1:

- A collection of HTML *posts*, each post seen as a unit of content, usually covering a single topic, possibly including comments added by readers, and being uniquely identified by a permanent URL (known as a *permalink*).
- A syndicated XML *feed*, comprising updates on the contents published in the blog, for easy access by client applications, known as aggregators. Two XML standards are in common use for blog feeds, namely Really Simple Syndication (RSS) [99] and Atom [166]. In addition, some blogs provide feeds for also retrieving comments.

[1] http://www.blogger.com.
[2] http://wordpress.com.
[3] http://smartdatacollective.com/matthewhurst/44748/farewell-blogpulse, accessed on January 14th, 2012.

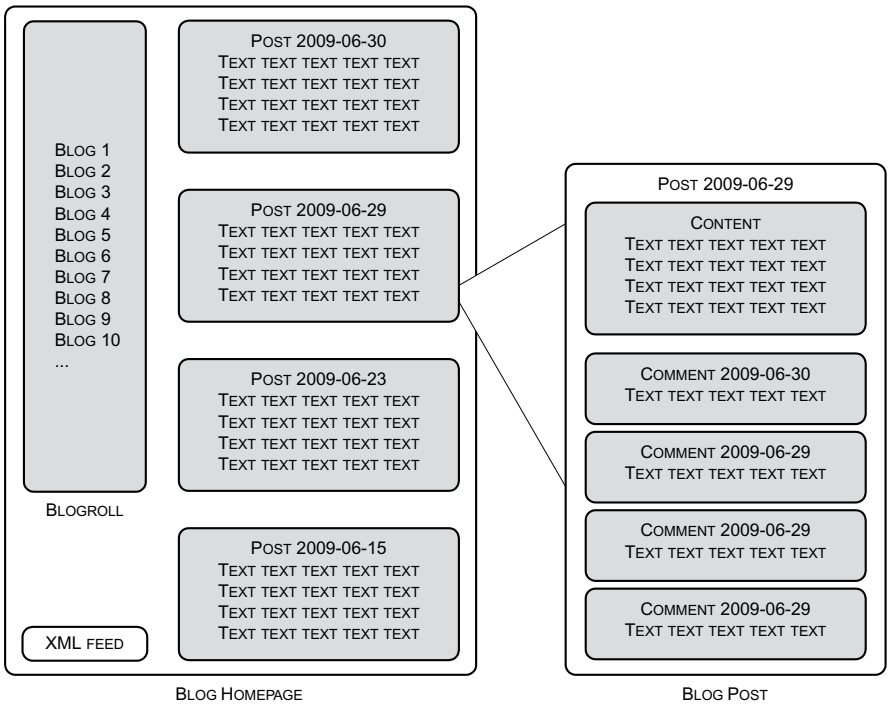

Fig. 1.1 Schematic view of a typical blog.

- An HTML *homepage*, with the latest posts in the blog organized in a reverse chronological order, and a list of "friend" blogs (i.e., those blogs that the blogger is interested in or is somehow related to), known as a *blogroll*.

Differently from traditional publishers, bloggers do not have to comply with strict guidelines regarding formatting or the use of formal language. Moreover, blog content is dynamic, in that it can be expanded, modified, or removed at any time. Besides text, blogs may include some multimedia content. In fact, there are blogs dedicated to publishing content of specific types — for instance, audio (podcasts), images (photoblogs), video (vlogs), etc. Recently, microblogs (e.g., Twitter) have also become popular as a means to publish very short content (e.g., a 140-character long post) about one's up-to-the-minute thoughts.

Indeed, Treem and Thomas [210] observed a common ambiguity in defining what a blog is. In a survey conducted with blog readers, no single defining attribute was identified as prevalent by the majority of the participants. "Commentary/opinion" was the most mentioned attribute (45%), followed by "thoughts/beliefs" and "diary/journal."

1.3 Why Do People Blog?

An important difference from the mainstream media is that blogs are regarded as "open gardens" [41], including by their authors. In other words, bloggers can bypass the control of the mainstream media in order to get their thoughts published and visible to a wide readership. Zhao et al. [240] recognised two types of bloggers: *specialists*, who write on specific topics, such as politics, technology, or sports, and many of whom receive thousands of visits every day on their blog; and *generalists*, who are typically ordinary people targeting much smaller audiences — in fact, many of their blogs function as personal diaries, reporting on the bloggers' daily activities.

Recent data [203] suggested an even balance between male and female bloggers, with 50.9% of bloggers being female, dispelling any notion of a gender divide among bloggers. Yet, a generation gap still exists, with only 7% of bloggers aged over 50. In contrast, over half of bloggers are aged 21–35, and 20% are aged 20 or under. Hence, teenagers form a significant percentage of the blogosphere, as well as many other social network communities. Their motivations were thoroughly examined by boyd [47], identifying the need for teenagers to "publicly" socialize, and the reduced availability of inter-personal communication in the digital era.

Oberlander and Nowson [168] classified blogger personalities along five classical dimensions: neuroticism, extroversion, openness, agreeableness, and conscientiousness. While extroverts would normally be more expected to blog, they found normal distributions for all of the dimensions except openness. Indeed, while bloggers were more likely to be open in nature, the observed traits of bloggers tended to follow those expected from other contexts. This showed how the act of blogging reflects rather than conceals the bloggers' personalities. For instance,

extroverts will document their life and emotions, neurotic bloggers act from an auto-therapeutic motivation, while blogs by open persons tend to contain commentary and evaluation [65].

Other attributes may be derivable from a blog other than from the writing style. For instance, Michelson and Macskassy [152] noted that a link to a Web site from a blog constitutes a consumption of that Web site. From that, inferences can be made, such as "has baby" or "has pet" with a reasonable degree of precision, but with low recall — e.g., the lack of a presence to a children's clothes shop Web site does not eliminate the fact that the blogger may have a young child.

In contrast to personal blogs, group blogs are of increasing popularity [84], where multiple authors can pool resources to create an interesting, coherent blog. One example of group blogging is corporate blogging. For instance, an internal blog within an organisation can enhance the communication among its employees; an external blog provides a more conversational public relations medium [165]. Indeed, even some traditional publishers, such as newspapers and other news outlets, have embraced blogging in face of the increasing competition.[4] Group blogs are in general more likely to be regarded of high quality, with higher link popularity and longer post lengths [84].

1.4 The Blogosphere

The rise of the *blogosphere* — the collection of all blogs on the Web — has changed not only the way information is consumed online but, more importantly, the way it is produced. Instead of being managed by a central entity, the content on the blogosphere is produced by millions of independent bloggers, who can write about virtually anything. The major difference from traditional publishers, however, is that blogs enable interaction. Interested readers can follow the published content regularly, or even subscribe to a blog's syndicated feed in order to automatically receive notifications of updates. More importantly, readers can comment on blog posts, hence effectively engaging in a discussion with the blogger and the other commentators [157] — in

[4] For instance, see http://blogs.guardian.co.uk or http://www.bbc.co.uk/blogs/.

fact, as bloggers are usually themselves readers of other blogs, the roles of information producer and consumer are often interchanged. Moreover, commenting plays a fundamental aspect in the popularity of a blog [226].

Another important form of interaction in the blogosphere is linking. Apart from "comment links", i.e., traces of commenting actions manifested as hyperlinks, inter-blog links can be roughly categorised into three main classes: blogroll links, citation links, and linkbacks. Blogroll links are usually placed on a blog homepage and point to "friend" bloggers — this relationship, however, does not necessarily correspond to a real-world friendship tie [5]. A citation link is similar to informational hyperlinks present in general Web pages in that it conveys the author's testament that the linked blog (or blog post) is somehow relevant to the context in which the citation is made. Finally, a linkback — also known as a trackback in its most popular variant — is a special mechanism that allows bloggers to keep track of who is linking to their posts. Together, these different forms of interaction help grow the blogosphere as a network of interconnected bloggers.

In aggregation, the perspectives of individual bloggers on a subject matter help elicit the public sentiment — the so-called "wisdom of the crowds" [202] — about this matter. Indeed, the blogosphere responds to real-world — perhaps newsworthy — events in a "bursty" fashion [109]. Gruhl et al. [74] characterised the diffusion of information on the blogosphere as consisting of long-running "chatter" topics, formed by "spike" topics generated by outside world events or, occasionally, "resonances within the community." Adamic and Glance [1] examined the U.S. political blogosphere during the 2004 presidential elections, and found the linkage behavior within the community of conservative blogs to be denser than that in the liberal community.

Any open Internet communication medium will be targeted by adversarial usage, often in the form of spam. In the blogosphere, several forms of spam have been observed, each driven by the easy accessibility of the technology: *spam blogs (splogs)* are blogs with fake content created with many hyperlinks, to increase the search rankings of other affiliated Web sites, as a form of "black-hat" search engine optimization (SEO); *fake blogs* are also blogs created for nefarious purposes, this

time where content is copied from bona fide blogs using their RSS feeds, then published, to attempt to gain revenue from ads hosted on the fake blog; *comment spam*, where bots publish comments on blog posts containing links for SEO purposes [155]; similarly, *trackback spam* takes advantage of common blog APIs that allow incoming links to a blog post to be shown on the original post, to create fake links to Web sites.

As alternative networked communities such as Facebook and Twitter have risen, the blogosphere has become increasingly interconnected with them. Indeed, 87% of bloggers have a Facebook account [197]. Such networks are self-reinforcing: a user may follow the tweets of a blogger that they read; links from tweets, Facebook updates, or LinkedIn posts drive a great deal of the incoming traffic to blogs [197].

1.5 Search on the Blogosphere

The advent of blogging as a publishing paradigm has led to an increasing mass of content being produced collectively by millions of bloggers worldwide, making the search for trustworthy, high-quality information on the blogosphere a challenging task. Indeed, Cho and Tomkins [41] identified issues for why search on social media such as the blogosphere is challenging: vulnerability to spam (facilitated by the ease that users can create content); short lifespan (public interest in a "hot" topic subsides rapidly over time); and locality of interest (with traditional media, content creation and publishing costs means that published content is intended to be of widespread interest, while a teenager's blog may only be of interest to his direct family and friends).

Similarly to traditional search tasks, blog search tasks can also be classified as adhoc or filtering [15]. In a typical adhoc search task, users submit different queries to a relatively static document collection.[5] A common instantiation of *ad hoc* search on the blogosphere is the search for blog posts that are relevant to the topic of the query. Additionally, motivated by the opinionated nature of blogs, this task can be enriched by considering posts that express a clear (positive or negative) opinion about the topic of the query. A filtering task, on the other hand,

[5] In the case of Web search engines, a static snapshot of their indices.

is characterised by documents being continuously retrieved against a fixed user query, as they are added to the collection. This task forms a popular usage of blog search engines [156], with users subscribing to updates from the content exposed by blogs in the form of syndicated feeds. The key challenge here is to identify high-quality blogs (e.g., from authoritative bloggers) that are worth following.

Thelwall [207] highlighted the benefits of searching the blogosphere from a social science perspectives. In particular, he pointed out that blog search engines facilitate the analysis of the public opinion about a particular subject, e.g., by analyzing the volume of posting activity relating to the subject over time, or by providing access to blog posts about the subject at a given point in time. Nevertheless, the observed trends are naturally only representative of the population of bloggers and do not necessarily represent the general population.

Overall, the blogosphere offers a challenging environment for creating effective search engines, characterized by its dynamic nature, the inherent structure, and how it responds and resonates to internal and external events. In the past decade, a great deal of research has addressed various points dealing with search on the blogosphere. In this survey, we aim to provide an overview of much of this research.

1.6 Scope of this Survey

This survey focuses on approaches to various search tasks, primarily those evaluated on publicly available blog corpora, such as the ones created in the context of the Blog track of the Text REtrieval Conference (TREC) [134, 136, 171, 173, 174] and the ICWSM Data Challenges. Additionally, we cover search tasks that are not necessarily targeted at the blogosphere, but that still leverage information from blogs as a means to enable other search tasks. Lastly, we discuss open directions in the field of blog search, and provide an introduction to the emerging field of search on microblogging environments. Outside the scope of this survey are approaches that use the blogosphere for tasks other than search (e.g., pure sentiment analysis), for which there are already excellent surveys (e.g., [178]).

Table 1.1. Notations used in this survey.

Notation	Definition
elements	
q	A user query
t	A unigram (e.g., a term or a term feature)
υ	An n-gram (e.g., a compound, passage or sentence)
p	A blog post
b	A blog
d	A day of interest
s	A news story
sets	
$\mathcal{C}$	A corpus of items (e.g., blog posts, blogs, news stories)
$\mathcal{L}$	A lexicon of terms
$\mathcal{Q}$	A set of queries
$\mathcal{D}$	A set of retrieved items
$\mathcal{F}$	A set of feedback items
$\mathcal{R}$	A set of relevant items
$\mathcal{O}$	A set of relevant and opinionated items
operators	
Γ_x	The set of lines in x
Υ_x	The set of n-grams in x
n_x	The cardinality of x
l_x	The length of x
df_x	The number of items (e.g., blog posts, blogs) where x occurs
sf_x	The number of n-grams where x occurs
$tf_{x,y}$	The number of occurrences of x in y
$pf_{\langle x_1,x_2\rangle,y}$	The number of occurrences of the pair $\langle x_1,x_2\rangle$ in y

When describing approaches to different blog search tasks, we will rely mostly on the notations described in Table 1.1.

The remainder of this survey contains the following:

- Section 2 discusses the history of information retrieval for blogs and the information needs on the blogosphere.
- Section 3 discusses approaches for searching for blog posts.
- Section 4 presents approaches for searching for entire blogs.
- Section 5 discusses how the blogosphere can aid other search tasks, such as identifying newsworthy or trendy topics.
- Section 6 describes publicly available resources that can aid research on blog search tasks.
- Section 7 discusses ongoing and open research directions on searching the blogosphere and other social media channels.

2

Blog Information Retrieval

The open, social, and rapidly evolving nature of the blogosphere has created a need for organizing and making available the vast amount of information published by bloggers every day. In this section, we draw parallels between how blogs have been organized and the wider history of searching the Web. Additionally, we describe existing blog search engines and discuss the information needs of blog searchers.

2.1 Blog Directories

As the blogosphere grew, bloggers and their readership found insufficient tools to allow them to find other blogs to subscribe to. Indeed, 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 blogs that they read or endorse. Readers of their blog may then use this list to find other interesting blogs.

However, the use of blogrolls drive the information discovery process for the reading user in the manner provided by the bloggers, and do not permit the reader to find blogs that are not mentioned by the blogs they are currently reading [90]. To counteract this, there exist a number

of blog directories — such as Blogflux[1] and Topblogarea,[2] to name but a few — containing submissions of blogs, which users can use to find blogs of interest. Figure 2.1 shows an example listing by the Blogflux directory for the politics topic. Glance et al. [68] discussed other blog directories or "census" sites that existed in 2004.

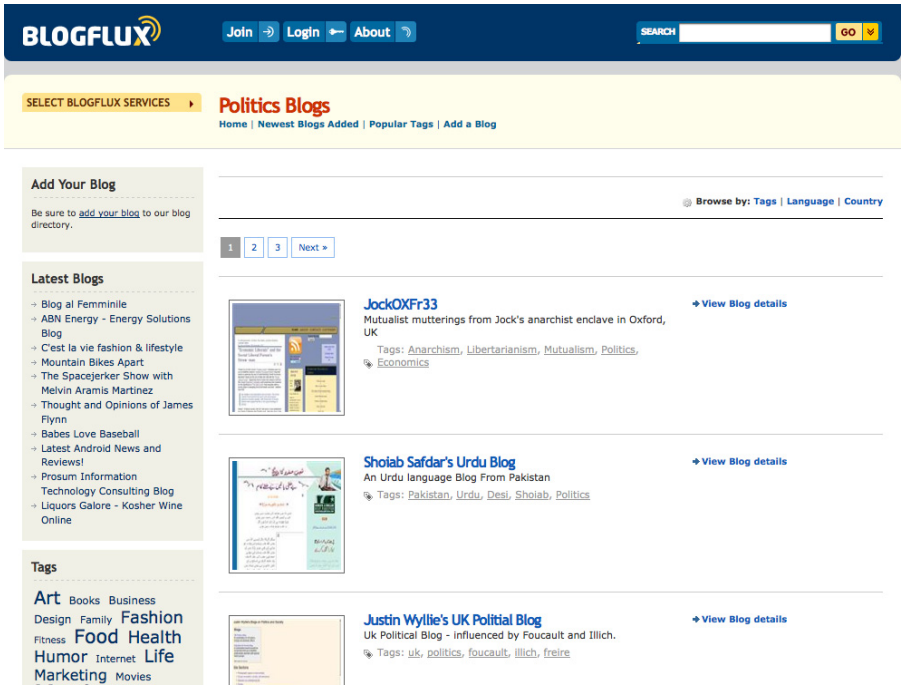

Fig. 2.1 Blogflux.com, an example of a blog directory.

The presence of blog directories is to some extent reminiscent of the early history of Web search in general. In particular, while early search engines such as Altavista, Excite, or Inktomi existed in the 1990s, many users resorted to Yahoo!'s manually catalogued ontology of links. Only once improved search engines were available (such as Google), did directories reduce in prevalence and popularity (Yahoo! built a search engine on top of its directory).

[1] http://www.blogflux.com/.
[2] http://www.topblogarea.com.

2.2 Existing Blog Search Engines

The rise in popularity of blogs was closely followed by the rise of blog search engines, helped by the easy monitoring of updates using the RSS feeds that each blog provides. Initially, blog search engines only provided blog post search, i.e., a list of relevant blog posts returned in response to a user query. However, as discussed in the previous section, the presence of blogrolls and the various manually categorized blog directories suggest that there is an underlying user task that blog post search does not cater for, namely, the search for entire blogs.

The now defunct DayPop is reported as one of the earliest blog search engines [208]. It integrated blog and news feeds to provide access to the most recent blog posts about the query terms. Indeed, freshness and recency is a key component of many deployed blog search engines, with some going as far as ranking posts in reverse chronological ordering [154, 208]. Indeed, as we will see in the next section, many queries to blog search engines concern recent real-world events, hence focusing the search engine on the most recent blog posts is a key aspect.

Thelwall and Hasler [208] conducted a survey of the available blog search engines, basing their analysis upon several technical aspects: support for advanced query language, coverage of common and rare terms, as well as the same term in different languages, consistency of results over time, spam, overlap between search engines, and precision of top-retrieved documents in terms of matching the terms of the query. However, their evaluation did not measure the actual relevance or usefulness of the retrieved documents for the query, as would be handled by a classical information retrieval effectiveness evaluation.

Technorati[3] is regarded as one of the definitive blog search engines, having been around since 2002. In contrast, Google waited until September 2005 to launch their blog search product.[4] Figures 2.2 and 2.3 show blog search results for the query "higgs boson" entered into both Technorati and Google's blog search product. In both interfaces, we can see support for two modes of blog search: searching for posts; and searching for blogs (called "blog homepages" in the Google

[3] http://technorati.com.
[4] http://www.google.com/blogsearch.

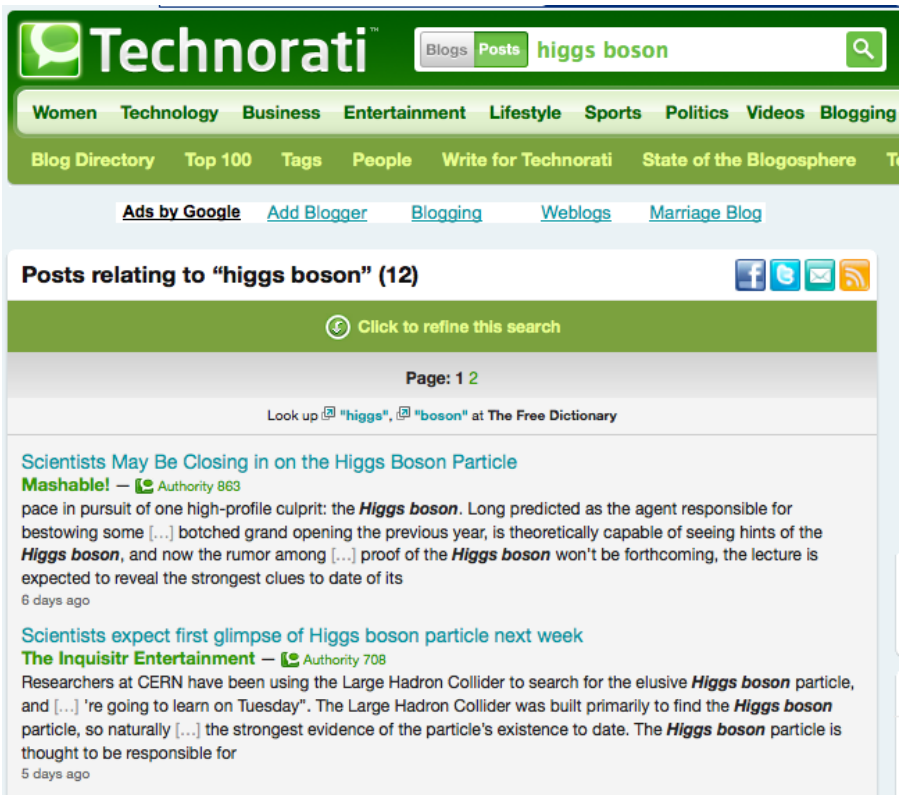

Fig. 2.2 An example of a Technorati post search results page for "higgs boson."

interface). In particular, Google places related blog results above the post ranking for this query, but does not do so for all queries. This suggests that it attempts to differentiate between queries with and without blog search intents. In contrast, Technorati does not combine the search results, instead providing a slider switch interface element to allow users to tell the search engine whether they are looking for blogs or posts.

Turning to the post ranking, we also note differences in the recency of the search results for the query, with Google showing more fresh results than Technorati in the initial "relevance" ranking. In contrast, Technorati gives default importance to the deemed authority of the ranked blogs, trying to focus on popular blogs. Both interfaces identify

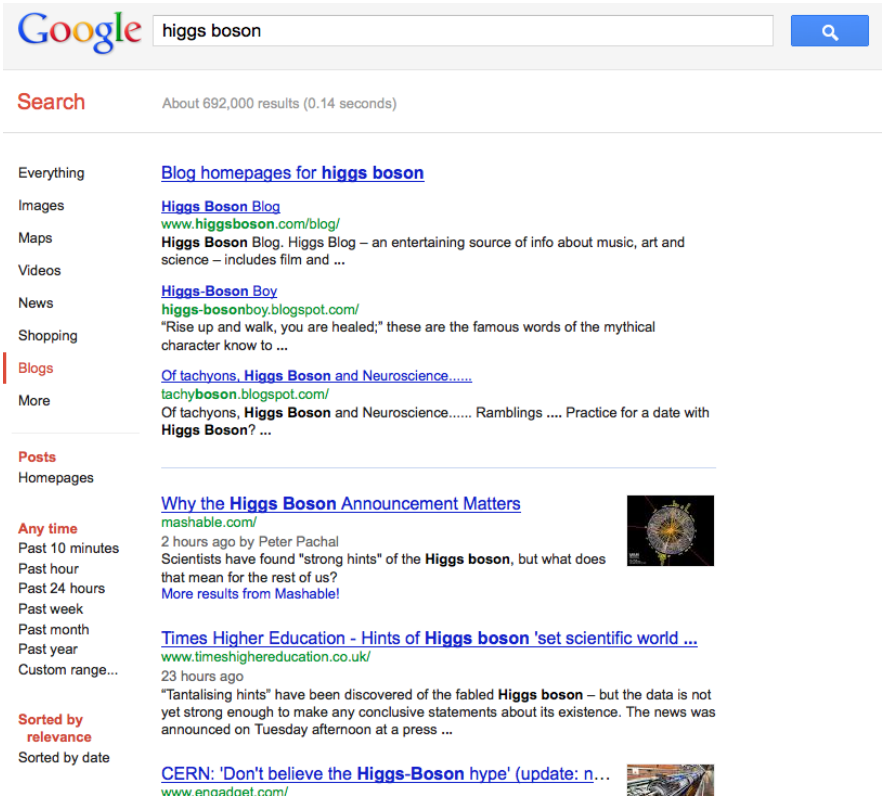

Fig. 2.3 An example of a Google blog search results page for "higgs boson" — facets to aid the user to "drill down" into the results are on the lower left-hand side of the results page.

that the user may have complex information needs, and wish to explore the results in different ways (e.g., reverse chronological or relevance ordered rankings), or over different time periods. In particular, Google provides this as a faceted search interface, with refinement options displayed down the left-hand side of the results page [81].

A different type of blog search engine addresses the monitoring of the blogosphere from a temporal perspective. For instance, as the blogosphere is full of discussion and subjective content, it can be useful to mine for insights and opinions about a company or product. Thelwall and Hasler [208] refer to this mechanism as a *graph search* engine, where the search engine is used to plot the number of mentions, or the polarity of opinion about the query terms.

BlogPulse [68] is one such graph search engine, which also provides blog post search. Figure 2.4 shows an example of its trending output, comparing the "buzz" on the blogosphere for three physics-related topics in late December 2011. We can see that blogs discussing CERN increased in frequency due to two different real-world events, namely reporting of observations of faster-than-light neutrinos, and the announcement of results concerning the Higgs Boson. BlogScope [21] is a similar search engine, which provides insights into current events, by displaying recent blog posts, along with trend graphs and temporally related query terms.

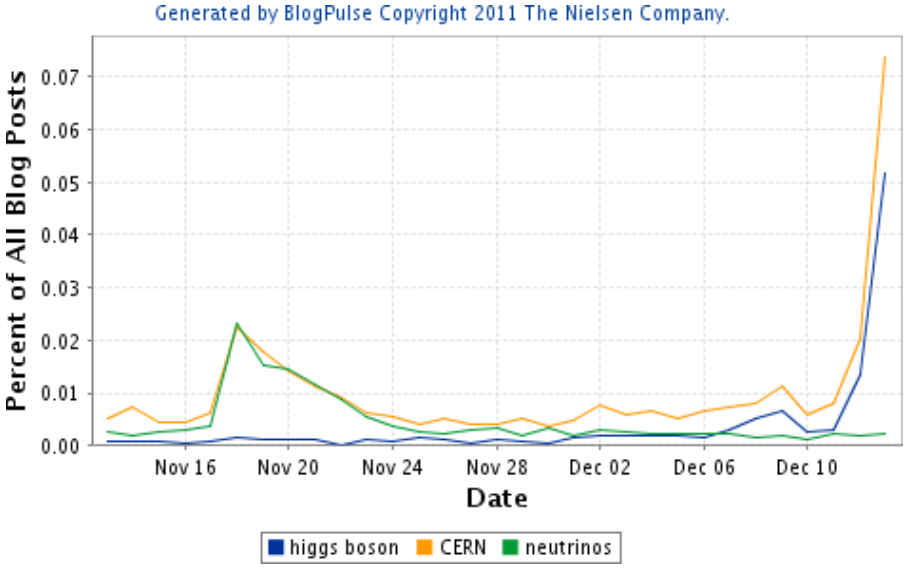

Fig. 2.4 An example of a BlogPulse trend graph for three physics-related queries.

Other companies address this information need from a purely commercial perspective, where clients pay for full round social media measures and advice. For instance, Cymfony[5] provide enterprise tools that permit clients insights into the voices of consumers within social media. Similarly, Clarabridge[6] allows clients to compare the social perspectives and discussions about their product or brand with

[5] http://www.cymfony.com.
[6] http://www.clarabridge.com/.

competitors. Converseon[7] can provide "social reputation scorecards" for companies, complete with comparisons to competitors.

In the remainder of this section, we describe a categorization of the information needs underlying all the previously described blog search tasks, with a view toward establishing the motivations for the blog search approaches described in the subsequent sections.

2.3 Information Needs on the Blogosphere

According to the influential work by Broder [28], Web searchers' information needs are traditionally classified into three main categories: *informational* ("find information on a particular topic"), *navigational* ("find a specific Web site"), and *transactional* ("perform a Web-mediated activity"). Mishne and de Rijke [156] proposed to extend this taxonomy to categorize the information needs on the blogosphere. Claiming that transactional queries are not naturally applicable to the blogosphere and that navigational queries can be answered by a general-purpose search engine (not necessarily a blog search engine), they investigated possible refinements to the widely accepted definition of informational queries in the context of blog search tasks.

In their study, the full query log of Blogdigger.com — a former second-tier blog search engine — as of May, 2005, was analyzed. This query log comprised 1,245,903 queries, 166,299 of which were unique. Filtering queries corresponded to most of the volume in the query log (81%), as they tend to be repeatedly submitted to the search engine in order to retrieve updated results (e.g., the user subscribes to the RSS results of a given query within their RSS reader). Among the unique subset of queries, however, *ad hoc* queries were predominant (70%).

By manually classifying two subsets of queries sampled from this query log, they derived important observations. Firstly, they analyzed a random set of 1,000 queries evenly distributed between *ad hoc* and filtering queries, denoted Random in Figure 2.5. In this subset, which represents the long-tail of the query log, they observed a predominance

[7] http://converseon.com.

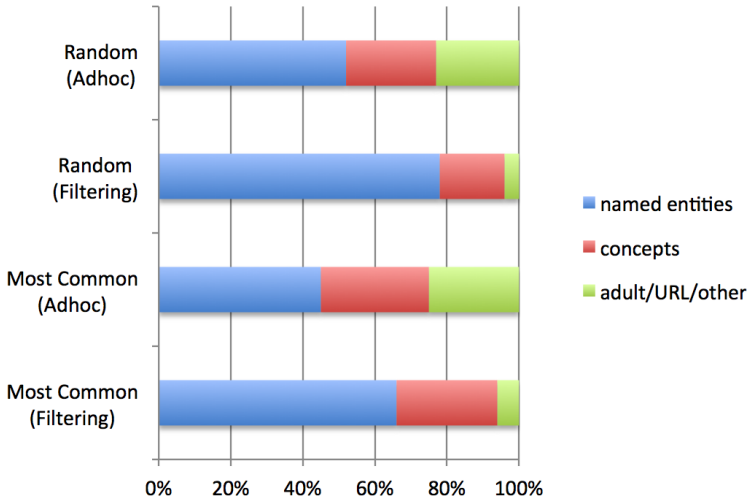

Fig. 2.5 Distribution of types of queries found within a blog search query log by Mishne and de Rijke [156].

of named entities (52% of *ad hoc* queries, and 78% of filtering queries), such as names of people, organizations, products, or locations. Among the remaining queries, most (25% and 18% of the *ad hoc* and filtering queries, respectively) were related to broader concepts, such as "stock trading" or "humor." The remainder of the queries were mostly adult-oriented or navigational queries. Additionally, a second subset with the 400 most popular queries in this query log — the head of the distribution — was also analyzed. In this subset, denoted Most Common in Figure 2.5, a similar trend as in the random set was observed, however with a different distribution, with a stronger emphasis on broader concepts — and lesser on named entities, although the latter was still predominant — when compared to the random set. An exception was adult-oriented queries, which were substantially more common in the subset of popular queries than in the random subset. Another fairly frequent query type in the subset of popular queries (20% of *ad hoc* and 15% of filtering queries) was that of news-oriented queries, suggesting that blog searchers are also interested in the blogosphere response to real-world events reported by the mainstream media. Based on these observations, Mishne and de Rijke [156] proposed a classification of blog

information needs into two broader categories — again, refinements of the higher-level *informational* category as defined by Broder [28]:

Context Queries. These queries are intended to identify the context around the occurrences of a named entity ("find what people say about a given entity"). These include well-known named entities (e.g., politicians, celebrities, major brands or events) as well as less known ones (e.g., people the searcher is closely related to, or the company the searcher works for).

Concept Queries. Queries of this type are intended to find blogs or blog posts related to the searchers' general interests (e.g., music, sports, politics) or authored by bloggers from a particular geographic region (e.g., the searcher's home town, or from a place the searcher wants to know about).

These types of queries have a direct correspondence with the two forms of information needs noted in the previous section: for context queries, the users have a well defined event about which they want information or discussion (probably recent), and this is likely to manifest with the searcher looking for blog *posts* to read; in contrast, for concept queries, the searcher has a broader information need, and is likely to be interested in blogs that they can subscribe to. Note that since Blogdigger.com did not support the mining of trends (graph search), these queries are not represented in the categorization presented above [156].

The identification of blogs to read was given further treatment by Hearst et al. [82], who suggested that the searcher may wish to further specify the types of blogs they are interested in reading. Such dimensions could be presented in a faceted search interface [81], such that the refinement could be achieved through exploration of the retrieved blog search results. An example of such interface is shown in Figure 2.3 for Google's blog search product where, for instance, users can change the time range of the retrieved blog posts. Indeed, several possible facets of interest were identified by Hearst et al., including quality characteristics (original content versus commentary on other content), style, personality or tone (witty, serious, scientific, artsy, religious, business-oriented), and level of interaction (who are interacting, who is linking to the blog,

reputation of the blog's authors). This observation was key in inspiring research into faceted blog search retrieval models, which are further discussed in Section 4.5.

Lastly, Mishne [154] recognized the importance of "better ranking mechanisms," and linked improved blog search engines with the development of test collections along classical information retrieval lines, with improved blog search techniques. In the remainder of this survey, we address retrieval approaches for three different types of search tasks: Section 3 discusses approaches for identifying blog posts, Section 4 discusses approaches for identifying blogs, and Section 5 discusses search tasks that are aided by evidence mined from the blogosphere. Many of the described approaches are built on resources or evaluated using test collections that are described in Section 6, such as those of the Blog track of the Text REtrieval Conference [134, 136, 171, 173, 174].

3

Blog Post Search

A blog *post* can be seen as the blogosphere's equivalent for the notion of a *document* in traditional textual corpora. Indeed, different from an entire blog, which comprises posts covering multiple (yet typically related) topics, an individual blog post is usually made around a primary topic, and often contains on-topic comments or references to other on-topic content. As such, it arguably represents the most natural information unit for search on the blogosphere. In this section, we survey published approaches that tackle two distinct blog post search tasks. In particular, Section 3.1 describes approaches for finding *relevant* blog posts, whereas Section 3.2 surveys approaches dedicated to finding *opinionated* blog posts in response to a user's query. Many of the described approaches have been tested on publicly available blog datasets provided by the Blog track of the Text REtrieval Conference (TREC) or the data challenge track of the Conference on Weblogs and Social Media (ICWSM), which will be covered in Section 6.

3.1 Ad hoc Search

An *ad hoc search* task models a scenario where users issue different queries to a static corpus — or a static snapshot of a dynamic

corpus — about a given topic. For each query, the goal of a retrieval system is to return as many relevant items as possible [15]. In the context of the blogosphere, an *ad hoc* search task can hence be stated as "*find me blog posts about* x," where x is the topic of interest.

The blogosphere poses additional difficulties for *ad hoc* search, which are related primarily to the presence of noisy and adversarial content. In the former case, a text cleaning procedure can be performed at indexing time, or improved term statistics can be leveraged from external, high-quality corpora at retrieval time. In the latter case, spam detection techniques are usually employed. In the remainder of this section, we describe approaches that tackle these different problems in order to improve *ad hoc* search on the blogosphere.

3.1.1 Text Cleaning

Blog search datasets typically include the raw content crawled from the blogosphere. In the case of blog posts, in particular, this usually comprises irrelevant content besides the core post topic and its associated comments, such as advertisements, client-side scripting code, and frequently a whole HTML template structure [171]. Such a noisy content can hurt not only the effectiveness of retrieval systems, but also their efficiency. Indeed, the noisy content may represent a significant fraction of the actual size of each blog post, increasing the demand for storage space and the total indexing time.

In order to remove this noisy content and extract cleaner content for indexing, "boilerplate removal" algorithms can be applied. Many such algorithms have originally been developed for general Web documents [36, 56, 215], while others have been developed specifically for the blogosphere. For instance, Nam et al. [164] proposed the DiffPost algorithm which removes irrelevant content from blog posts by assuming that posts from the same blog follow the same HTML template. As a result, they consider common parts of different posts from the same blog as representing irrelevant content. More formally, let Γ_p be the set of lines in post p. Given a reference blog post p_r, the relevant content of a target post p_t from the same blog as p_r is defined as $\Gamma_{p_t} \backslash \Gamma_{p_r}$, whereas its non-relevant content is defined as $\Gamma_{p_t} \cap \Gamma_{p_r}$. This simple

heuristic showed a very high precision (above 98%) at identifying noisy portions of blogs posts. More importantly, by applying DiffPost, significant improvements were observed in retrieval performance [115], along with a reduction of 40% in content size [164].

Kohlschütter et al. [102] also addressed the removal of boilerplate within the blogosphere. In particular, after splitting a blog post into individual HTML elements, each element was classified as boilerplate or not, based on shallow textual features. They found that the number of words and the link density were accurate features. Once again, significant improvements in the effectiveness of a blog search engine were observed after boilerplate removal.

3.1.2 Spam Detection

Similarly to the general Web, the blogosphere is also severely affected by spam [104]. A spam blog (or *splog*) typically hosts blog posts automatically created with content plagiarized from legitimate sites, so as to gather audience and increase revenue with ads, or to serve as a link farm for boosting the authority of other affiliated sites [33].

The presence of such deceptive content can degrade the performance of blog search systems. In particular, Macdonald et al. [135] analyzed the blog post rankings produced by the participant systems at the TREC Blog track from 2006 to 2008 [134, 171, 173], and found that around 10% of all retrieved posts were actually spam. While removing these spam posts did not alter the relative retrieval performance of the considered systems, their absolute performance was generally improved.

Splogs represent a challenging obstacle to effective retrieval due both to their volume and to the difficulty in their detection. Indeed, using a crawl of the blogosphere provided by the BlogPulse[1] search engine, Kolari et al. [104] found that around 75% of all content update notifications received by ping servers — which maintain a list of blogs with new content — comes from spam blogs. In order to automatically detect spam on the blogosphere, they later proposed to integrate both content and link-based features using Support Vector Machines (SVM) [103, 105]. Content features included n-gram frequency on the

[1] http://blogpulse.com.

post content, outgoing anchor-text and URLs, as well as the presence of named entities and the post compression ratio. Link features were based on the likelihood of a blog linking to a (known) splog. While unigram features showed a very high discriminating power, the integration of link-based features did not show further improvements.

A related approach for classifying spam blogs using SVM was proposed by Lin et al. [123]. Different from other approaches, however, they proposed to exploit the dynamic nature of the blogosphere. In particular, they developed features to capture the self-similarity of a blog over a time period in terms of its posting frequency, content, and linkage. These features were then combined with standard content features computed from different parts of a blog, including its title and homepage, as well as its posts' title, content, and anchor-text. An evaluation based on 1,600 manually labelled blogs from the TREC Blogs06 corpus showed that integrating the proposed temporal features can significantly improve over using the content-based features alone.

A common issue with supervised approaches for spam classification is that their performance tends to degrade over time, as spammers begin to explore new and more sophisticated tactics for deceiving retrieval systems. At the same time, continuously obtaining new training examples to periodically re-train a classification approach is expensive. To address this issue, Chinavle et al. [40] proposed a method that builds upon not one, but an ensemble of diverse classifiers. In particular, they found that changes in the mutual agreement of different classifiers over time are indicative of decreased classification performance. By using this information to automatically trigger the re-training process, they showed that the ensemble can be maintained as accurate over time as if it was retrained frequently, albeit with significantly less computation.

Finally, a related problem to spam blogs is comment spam. In this modality, instead of automatically creating entire spam blogs, a spammer posts comments to genuine blogs selected at random, with links to pages affiliated with the spammer. This problem poses additional difficulties to retrieval systems that make use of authority features, as the spam-affiliated pages appear to be linked by legitimate (and sometimes authoritative) bloggers. To address this problem, Mishne et al. [155] proposed an unsupervised approach by comparing the language model

of a blog post to those of its comments, and those of the pages linked to by each comment. The more the language model of a comment (or its linked pages) diverges from that of the blog post, the more likely the comment is spam. Initial results on a small dataset comprising a random crawl of the blogosphere showed that their approach substantially outperforms a naive classifier using maximum likelihood estimation.

3.1.3 Relevance Estimation

A strong *ad hoc* search performance is usually the basis for performing well in more involved tasks, such as the opinion search task discussed in Section 3.2. However, accurately estimating the relevance of a blog post for a query is in itself a challenging task. Not only do blog posts include noisy and adversarial content, as described in the previous sections, but they are also usually written in an informal, subjective language, mixing characteristics of written and spoken languages [154]. To overcome the vocabulary mismatch between query terms and blog post content, several approaches have been proposed in recent years.

3.1.3.1 Passage Retrieval

Lee et al. [115] experimented with a passage retrieval approach for *ad hoc* search in blog posts. Their intuition was that a blog post may contain many off-topic passages, which should not contribute for the post to be retrieved. Accordingly, given a user's query q, they scored each blog post p by linearly combining its relevance score with the score of its highest scoring passage $\upsilon \in \Upsilon_p$:

$$score(q,p) = (1-\alpha)\,\mathrm{P}(q|\theta_p) + \alpha \max_{\upsilon \in \Upsilon_p} \mathrm{P}(q|\theta_\upsilon), \tag{3.1}$$

where θ_p and θ_υ are the post and passage language models, respectively. In particular, $\mathrm{P}(q|\theta_p)$ and $\mathrm{P}(q|\theta_\upsilon)$ were estimated using maximum log-likelihood estimates with Dirichlet and Jelinek–Mercer smoothing [231], respectively, with the parameter α controlling the mixture of the combination. Their approach assumed passages of an arbitrary size [160]. Even under this simplifying assumption, it was shown to deliver one of the top *ad hoc* search performances in the TREC 2008 Blog track [173].

On a similar vein, Chenlo and Losada [37] investigated the use of sentence-level features to improve *ad hoc* search in blog posts. To this end, they scored every sentence υ from the top blog posts retrieved by BM25 [185] using a sentence-level vector space approach [6]:

$$score_{\upsilon}(q,\upsilon) = \sum_{t \in q} \log(tf_{t,q}+1) \log(tf_{t,\upsilon}+1) \log\left(\frac{|\Upsilon_{\mathcal{C}}|+1}{0.5+sf_t}\right), \quad (3.2)$$

where $tf_{t,q}$ and $tf_{t,\upsilon}$ are the number of occurrences of the term t in the query q and sentence υ, respectively, $|\Upsilon_{\mathcal{C}}|$ is the total number of sentences in the collection $\mathcal{C}$, and sf_t is the number of such sentences where the term t appears. To produce a post-level score, they aggregated the scores of all sentences $\upsilon \in \Upsilon_p$ for each post p:

$$score_{\upsilon}(q,p) = \Psi_{\upsilon \in \Upsilon_p}\, score_{\upsilon}(q,\upsilon), \quad (3.3)$$

where Ψ is a summary statistic, such as the number of high-scoring sentences, as well as the average, variance, median, and maximum sentence score in each post. This post-level score was then linearly combined with the baseline $score_r(q,p)$, given by BM25 [185], as follows:

$$score(q,p) = \alpha\, score_r(q,p) + \beta\, score_{\upsilon}(q,p), \quad (3.4)$$

where α and β are free parameters. Using the TREC 2007 and 2008 Blog track queries [134, 173], they found that the number of high-scoring sentences and the median sentence score improved significantly compared to BM25. While the other features did not show significant improvements, the authors hypothesized that a multi-feature ranking approach could lead to further improvements.

3.1.3.2 Concept Matching

Another effective approach for *ad hoc* search in blog posts exploits noun phrases in a user's query. Such an approach is particularly motivated by the high volume of queries with named entities in blog search engines [156] — the so-called context queries described in Section 2.3. For instance, Jia et al. [92, 237] proposed to identify concepts in the query, including proper nouns, dictionary phrases, simple phrases, and complex phrases. Proper nouns comprised named entities such as

people, locations, and organizations, primarily derived from Wikipedia, while dictionary phrases were mostly extracted from WordNet.[2] Any other grammatically valid n-grams were identified as either simple (for bigrams) or complex (for n-grams of higher order) phrases, using a probabilistic parser with statistics obtained from Google [235]. Given the query q with matched posts $\mathcal{D}_q$, a ranking π was produced by sorting the posts $p \in \mathcal{D}_q$ by their concept- and term-based scores, the latter serving as a tie-breaking criterion [127]:

$$\pi(q, \mathcal{D}_q) = sort_{\left\langle \begin{smallmatrix} score_{\upsilon}(q,p) \\ score_{r}(q,p) \end{smallmatrix} \right\rangle} \{p \in \mathcal{D}_q\}, \tag{3.5}$$

where the term-based $score_r(q,p)$ was estimated using BM25 [185] and the concept-based $score_{\upsilon}(q,p)$ was defined as:

$$score_{\upsilon}(q,p) = \sum_{\upsilon \in \Upsilon_q \cap \Upsilon_p} \left(idf_{\upsilon} + \sum_{c \in \upsilon} idf_c \right), \tag{3.6}$$

where υ is a concept in the intersection of the concept sets for the query q (i.e., Υ_q) and the post p (i.e., Υ_p), c is a possible simple concept in υ, and idf_{υ} and idf_c are the inverse document frequencies of υ and c. Notably, while scoring a blog post, possibly abbreviated forms of the query concepts in that particular post were also accounted for as valid concept occurrences. Therefore, their approach effectively exploited abbreviations as introduced by individual authors, as opposed to generalizing abbreviations produced by different authors [236].

An alternative strategy to identify multiple phrases in a query was proposed by Vechtomova [213]. In particular, her approach considered the longest subsequence of the query that matched a Wikipedia title as a phrase, by iteratively attempting to match gradually smaller subsequences [138]. Using the query set from the TREC 2008 Blog track [173], she showed that, by up-weighting the occurrence of phrases identified according to this simple heuristic, her approach significantly improved upon a baseline ranking produced by BM25 [185].

[2] http://wordnet.princeton.edu.

3.1.3.3 Query Expansion

Another effective approach commonly employed to enhance the topic-relevance performance of document retrieval systems in general is query expansion [186]. In its traditional application, the original query is submitted to the system and the top retrieved documents are used as the so-called pseudo-relevance feedback set. From this set, a term weighting model is then applied to select the highest weighted terms, which are finally added to the original query, forming the expanded query. In the context of *ad hoc* search in blog posts, several advanced query expansion strategies have been investigated. Similarly to their approach to first-pass retrieval, Lee et al. [115] experimented with passage retrieval for query expansion. In particular, instead of a pseudo-relevance feedback set comprising the top ranked blog posts for the query, they used the highest scoring passages in these posts, augmented with a fixed-length left and right context [159]. In order to identify highly scoring terms from these passages, they updated the query language model using the model-based feedback approach [230]. In combination with their passage retrieval approach, their passage-based feedback approach was shown to outperform all other approaches deployed in the context of the TREC 2008 Blog track [173] in terms of *ad hoc* search performance.

Analogously, Jia et al. [92, 237] also adapted their concept-based retrieval approach to improve query expansion. In particular, they proposed to expand a query using synonyms of the concepts identified in this query. To this end, their approach first attempted to find a Wikipedia entry page for each query concept. If such an entry page existed, its title was then added to the expanded query. Additionally, the expanded terms also included those that appear frequently in proximity to the original query terms in the Wikipedia entry page, as well as among Google search results, and the target collection of blog posts.

The use of external resources for query expansion — a technique generally known as collection enrichment — was also investigated by Weerkamp and de Rijke [220]. In particular, they proposed to build relevance models [113] using corpora external to the blogosphere. To form a feedback set for a query q, they selected the top 10 documents from an external resource. Besides Wikipedia, they also experimented

with a news resource temporally aligned with their target collection of blog posts. By doing so, they tried to exploit the connection between news and the blogosphere, in line with the observation that many blog search queries are news-oriented [156], as discussed in Section 2.3. From this feedback set, they selected the top 10 expansion terms e using Lavrenko's RM2 [113]. The final query model θ_q was then obtained by combining the expanded and original models, $\theta_q^{(e)}$ and $\theta_q^{(o)}$, respectively:

$$\mathrm{P}(t|\theta_q) = \lambda \mathrm{P}(t|\theta_q^{(e)}) + (1-\lambda)\mathrm{P}(t|\theta_q^{(o)}), \tag{3.7}$$

where the parameter λ controlled the mixture between the two models. Their approach was shown to deliver one of the top *ad hoc* search performances in the TREC 2007 and 2008 Blog tracks [53, 220].

Machine learning has also been used to improve query expansion for blog post *ad hoc* search. In particular, Zhang et al. [233] proposed a learning approach to automatically select effective expansion terms, i.e., terms that would result in an improved retrieval performance when added to the query. In their approach, from the top 120 posts retrieved for a given query q, an initial set of candidate terms was extracted using a standard Rocchio formulation [186]:

$$\vec{q} = \vec{q}^{(o)} + \alpha \sum_{p \in \mathcal{F}_q} \vec{p} \tag{3.8}$$

where $\vec{q}$ is the query vector resulting from the combination of the original vector $\vec{q}^{(o)}$ and the aggregate vector representing the terms in the posts p from the feedback set $\mathcal{F}_q$. In this formulation, the parameter α controls the weight given to pseudo-relevance feedback. The top 400 terms in the final query vector were then represented as feature vectors themselves, to enable a supervised selection of the most effective expansion terms. Term features comprised basically term and document frequency summaries (e.g., sum, average, max) across the feedback documents, whereas classification labels were defined as the observed improvement in average precision when adding the term to the initial query. Finally, a set of 200 potentially effective expansion terms was selected using an SVM classifier with a Radial Basis Function (RBF) kernel. This machine learning approach was shown to outperform an

unsupervised pseudo-relevance feedback baseline, and ranked among the top *ad hoc* search approaches in the TREC 2007 Blog track.

Overall, the *ad hoc* search approaches described in this section were shown to be effective to different extents. As a result, they provided the basis for effectively deploying techniques for more involved search tasks, such as opinion search in the blogosphere, as discussed next.

3.2 Opinion Search

Improving *ad hoc* search performance is desirable for retrieval systems in general. Nevertheless, search users are often engaged in more involved search tasks, which go beyond a simple search for relevant content. One such task, which has received considerable attention from the blog information retrieval community in recent years, is *opinion search*. Placed in the broader area of sentiment analysis [178], this task is motivated by the knowledge that can be extracted when bloggers are considered collectively. In other words, it aims to find out what the blogosphere is saying about a particular topic. Following the taxonomy described in Section 2.3, this task impersonates the information needs expressed mostly as context queries, and can be stated as "*find me blog posts with an expressed opinion about* x," where x represents a named entity.

Several approaches have been proposed in the past few years for the specific problem of identifying opinionated blog posts. Most of these approaches tackle the opinion search task as a re-ranking problem. In a first stage, their goal is to find as many relevant posts as possible, regardless of their opinionated nature. In a second stage, these posts are re-ranked using some opinion detection technique and an appropriate combination of scores. Amati et al. [11] deeply analyzed the interplay of topical relevance and opinion retrieval performance using Monte Carlo simulations. By randomly perturbing a system's performance at either the first or the second stage, their approach could assess, for instance, the minimum opinion detection accuracy required to improve over a particular *ad hoc* search baseline, or the best achievable opinion retrieval performance on top of this baseline. Over the years, it has been observed that a system's performance at the second stage is highly correlated with its performance at the first stage [134, 171, 173].

In this section, we focus on approaches devoted specifically to opinion search — i.e., mostly the second stage of the aforementioned two-stage approaches. The most effective of these approaches in the literature can be roughly organized into two main families: lexicon-based and classification-based approaches. In the following, we describe the most representative approaches in each of these families. Additionally, we describe polarity search approaches, specifically aimed at differentiating between positively and negatively inclined opinions.

3.2.1 Lexicon-based Approaches

Lexicon-based approaches build a list of terms with known semantic orientation, the so-called *opinionated lexicon* or *dictionary*. As described in Section 3.2.1.1, such approaches search for opinionated blog posts by quantifying the occurrence of lexicon terms in each post. Additionally, some approaches also exploit the strength of the relationship between the occurring lexicon terms and the topic of the query, so as to quantify opinionatedness in context, as described in Section 3.2.1.2. Lastly, in Section 3.2.1.3, we describe single-stage approaches to opinion search.

3.2.1.1 Context-Independent Approaches

Context-independent approaches aim to quantify the opinionatedness of a blog post regardless of the topic of a query. In particular, they assume that the posts highly ranked for this query are on-topic, and focus on estimating their likelihood of being opinionated.

For instance, Amati et al. [8] proposed an information-theoretic approach to automatically build an opinionated lexicon. In particular, they observed that opinionated terms possess low information content, in that they can hardly discriminate between relevant and non-relevant blog posts. On the other hand, such terms show a high discriminating power when differentiating the subset of opinionated posts from all relevant posts. To exploit this observation, they proposed to quantify the opinionatedness of a term t using the Kullback–Leibler (KL)

divergence:

$$w_t = KL(\vec{tf}_{t,\mathcal{O}} || \vec{tf}_{t,\mathcal{R}}), \tag{3.9}$$

where $\vec{tf}_{t,\mathcal{O}}$ and $\vec{tf}_{t,\mathcal{R}}$ represent the normalized frequency distributions of term t in the set $\mathcal{O}$ of blog posts assessed as opinionated and the set $\mathcal{R}$ of relevant but non-opinionated blog posts from the TREC 2006 Blog track [171]. This automatic selection introduced noisy terms, primarily due to frequency peaks in a restricted number of opinionated posts. To overcome this problem, they favored those terms that are more uniformly distributed in the set of opinionated blog posts, as these terms were more likely to convey an opinion regardless of a particular topic. In order to enable opinion search, the top terms in the obtained lexicon were submitted to a retrieval system as a query q_o, so as to assign each individual post p an opinionated score:

$$score_o(q,p) = \frac{score_r(q_o,p)}{rank_r(q,p)}, \tag{3.10}$$

where $score_r(q_o,p)$ is the score assigned by the Divergence From Randomness (DFR) DPH model [9] for the post p with respect to q_o, and $rank_r(q,p)$ is the rank position of p when scored by DPH with respect to the initial query q. This opinionated score was then integrated with the initial relevance score of p with respect to q, as follows:

$$score(q,p) = \frac{score_r(q,p)}{rank_o(q,p)}, \tag{3.11}$$

where the initial relevance score of p (i.e., $score_r(q,p)$) is modified by dividing it by the rank position of p according to its opinionated score (i.e., $score_o(q,p)$). Experiments using the TREC 2007 Blog track queries showed improvements of up to 19% (MAP 0.321) over the DPH relevance-only baseline (MAP 0.270) and 33% over the TREC 2007 median opinion retrieval performance (MAP 0.242) [134].

A similar approach was proposed by He et al. [79] to automatically derive an opinionated lexicon from the target collection itself, however with a different mechanism to weight a term's opinionatedness. In particular, from the list of all terms in the collection ranked by

their within-collection frequency (i.e., the number of posts in which the term occurs in the collection) in descending order, they first applied a skewed query model to filter out those terms that were too frequent or too rare [31]. This aimed to remove terms with too little or too specific information and which could not be interpreted as generalized, query-independent opinion indicators. The remaining terms from the list were then weighted using the TREC 2006 Blog track queries for training [171]. To this end, instead of simply computing the divergence of a term's frequency distribution between the sets of opinionated and non-opinionated posts, they employed a different term weighting model from the DFR framework [7]. In particular, the Bo1 model is based on the Bose–Einstein statistics given by the geometric distribution, which measures how *informative* a term is in the set $\mathcal{O}$ of opinionated posts in contrast to the set $\mathcal{R}$ of relevant but non-opinionated ones:

$$score_o(t) = tf_{t,\mathcal{O}} \log_2 \frac{1+\lambda_t}{\lambda_t} + \log_2(1+\lambda_t), \tag{3.12}$$

where $\lambda_t = \frac{tf_{t,\mathcal{R}}}{|\mathcal{R}|}$ is the mean of the (assumed) Poisson distribution of the term t in the relevant documents $\mathcal{R}$. The top weighted terms by Bo1 were then submitted as a query q_o and the retrieved posts scored using a traditional document ranking approach. In order to generate the final score for a post p, the score produced for this opinionated query was used as the opinion score of the post, $score_o(q_o, p)$, and was combined with the post's relevance score, $score_r(q, p)$, according to:

$$score(q,p) = \frac{-k}{\log_2 score_o(q_o,p)} + score_r(q,p), \tag{3.13}$$

where k is a free parameter. Both $score_o(q_o, p)$ and $score_r(q, p)$ were computed using the DFR InLB weighting model with term proximity enabled [7], with the former being further normalized to yield a probability distribution. Experiments using the TREC 2007 queries attested the effectiveness of the proposed approach (MAP 0.338), with improvements of up to 12% on top of the relevance-only baseline (MAP 0.303), and 40% over the TREC 2007 median (MAP 0.242) [134]. Moreover, a thorough analysis confirmed the superiority of Bo1 at weighting a term's opinionatedness in comparison to the KL divergence.

Another effective lexicon-based approach was proposed by Lee et al. [115, 161]. In their approach, an initial lexicon $\mathcal{L}$ was built using SentiWordNet [55], a publicly available lexical database for opinion mining. Each term $t \in \mathcal{L}$ was initially assigned the maximum weight $w_{t,SWN}$ across all of its possible senses $m \in \mu(t)$ in SentiWordNet:

$$w_{t,SWN} = \max_{m \in \mu(t)} \big[\max(\mathrm{P}(\mathcal{O}_-|m), \mathrm{P}(\mathcal{O}_+|m))\big], \tag{3.14}$$

where $\mathrm{P}(\mathcal{O}_-|m)$ and $\mathrm{P}(\mathcal{O}_+|m)$ denote the probability of the sense m being negatively or positively opinionated, respectively. To refine this general-purpose lexicon for the particular setting of blog post search, two weighting schemes were proposed: query-independent and query-dependent. In the *query-independent* scheme, each term in the lexicon was weighted by a mixture of two language models: an opinion model θ_o, built from product reviews, and a topical-relevance model θ_r, built from product specifications, both collected from Amazon.com:

$$w_t = \lambda_t \mathrm{P}(t|\theta_o) + (1 - \lambda_t)\mathrm{P}(t|\theta_r), \tag{3.15}$$

where the parameter $\lambda_t = w_{t,SWN}$ controlled the balance between the two generation probabilities. Their second term weighting scheme used pseudo-relevance feedback to assign terms a *query-dependent* opinion weight [161]. In particular, a query-dependent weight $w_{t,q}$ was computed by scoring each term $t \in \mathcal{L}$ according to the opinionatedness of the top-retrieved posts $\mathcal{F}_q$ where this term occurred:

$$w_{t,q} = \sum_{p \in \mathcal{F}_q} \mathrm{P}(\mathcal{O}|p)\,\mathrm{P}(p|t), \tag{3.16}$$

where $\mathrm{P}(p|t)$ was assumed to be uniformly distributed over the posts in $\mathcal{F}_q$ that contained the term t. In turn, the probability of p being opinionated was estimated as $\mathrm{P}(\mathcal{O}|p) \propto \sum_{t \in p} w_{t,SWN}\,/l_p$, where l_p denoted the length (in tokens) of post p. Alternatively, they experimented with using the best passage from each retrieved blog post (as opposed to the entire post) to form the pseudo-relevance feedback set $\mathcal{F}_q$ [160]. Using either the query-independent or the query-dependent lexicon, they quantified the opinionatedness of each blog post retrieved by a

relevance-only baseline and re-ranked these posts using a standard linear combination of relevance and opinion scores:

$$score(q,p) = (1-\alpha)\, score_r(q,p) + score_o(q,p), \tag{3.17}$$

where α controlled the balance between relevance and opinionatedness. Both $score_r(q,p)$ and $score_o(q,p)$ were given by BM25 [185], with the latter encapsulating $tf_{\mathcal{L},p}$ instead of the standard term frequency component, $tf_{t,p}$. Results using the TREC 2007 Blog track queries showed improvements of up to 17% (MAP 0.440) over the topic-relevance baseline (MAP 0.378), with a similar performance to the top performing approaches at the TREC 2007 Blog track (82% over the TREC 2007 median). On the TREC 2008 queries, this approach delivered the top performance (MAP 0.457) among the participant groups.

3.2.1.2 Context-Dependent Approaches

The aforementioned approaches explored the presence of opinionated terms in a blog post in order to improve opinion search performance. Nevertheless, the mere expression of opinion may be a weak signal, unless it is targeted at the topic of interest of a particular query. To address this, different approaches have been proposed to explicitly account for the occurrence of subjectivity in the context of the query.

One of the first approaches to exploit opinion in context was introduced by Santos et al. [190]. Their approach sought to promote retrieved blog posts where the query terms occurred near opinionated sentences. Besides the fact that a blog post containing opinionated sentences is more likely to be itself opinionated, their intuition was that such sentences provided a proper context for searching for opinions about the query terms. Accordingly, each blog post p retrieved for a query q was scored based on the occurrence of the query terms $t \in q$ in proximity to the (potentially opinionated) sentences $\upsilon \in \Upsilon_p$:

$$score(q,p) = \lambda_1\, score_r(q,p) + \lambda_2 \sum_{t \in q} \sum_{\upsilon \in \Upsilon_p} score_o(t,\upsilon), \tag{3.18}$$

where $score_r(q,p)$ was given by a standard relevance-only baseline. For the proximity-oriented $score_o(t,\upsilon)$, they adapted the DFR pBiL model,

which quantifies the co-occurrence of pairs of terms using the binomial randomness model [124]. In their adaptation, instead of a pair of terms, the pBiL model was used to quantify the co-occurrence of each query term t with each identified sentence υ in the post p:

$$score_o(t,\upsilon) = \frac{w_{t,q}\, w_{\upsilon,\mathcal{L}}}{pf_{\langle t,\upsilon\rangle,\Upsilon_p}+1}\Bigg[- \log_2(|\Upsilon_p|+1) \\ + \log_2(pf_{\langle t,\upsilon\rangle,\Upsilon_p}+1) \\ + \log_2(|\Upsilon_p| - pf_{\langle t,\upsilon\rangle,\Upsilon_p}+1) \\ - pf_{\langle t,\upsilon\rangle,\Upsilon_p} \log_2 \frac{1}{|\Upsilon_p|} \\ - (|\Upsilon_p|\, pf_{\langle t,\upsilon\rangle,\Upsilon_p}) \log_2 \left(1 - \frac{1}{|\Upsilon_p|}\right)\Bigg], \tag{3.19}$$

where $w_{t,q}$ is the weight originally assigned to the query term t, and $w_{\upsilon,\mathcal{L}}$ is the opinion weight of the sentence υ according to the lexicon $\mathcal{L}$, obtained by summing the opinion weights of all terms in the sentence. To compute opinionated term weights, Santos et al. employed two context-independent, lexicon-based approaches, leveraging either a manually [80] or an automatically built [79] lexicon. To quantify the co-occurrence of a term $t \in q$ and a candidate sentence $\upsilon \in \Upsilon_p$ in the post p, they computed the frequency $pf_{\langle t,\upsilon\rangle,\Upsilon_p}$ of the pair $\langle t,\upsilon\rangle$ among the set Υ_p of all sentences identified from p. Empirical results showed consistent improvements compared to five standard topic-relevance baselines provided in the context of the TREC 2008 Blog track [134], with gains of up to 69% (MAP 0.408) and 20% (MAP 0.396) over the TREC 2007 (MAP 0.242) and 2008 (MAP 0.329) medians, respectively.

A language modelling approach to exploit opinion in context was proposed by Seki and Uehara [194]. While traditional n-gram language modelling approaches can effectively capture short-distance term dependencies, they become less effective as wider contexts are considered, primarily due to the sparsity of higher-order n-grams [139]. To combat data sparsity while still exploiting the relationship between query terms and opinionated expressions, they proposed to leverage trigger language models [112]. Such models, originally developed

to capture long-distance term dependencies, estimate the probability $\mathrm{P}_\zeta(y|\mathcal{H}_y)$ of a term y being triggered by the terms $x \in \mathcal{H}_y$ that precede it:

$$\mathrm{P}_\zeta(y|\mathcal{H}_y) = \frac{1}{|\mathcal{H}_y|} \sum_{x \in \mathcal{H}_y} \alpha(y|x), \tag{3.20}$$

where x is denoted a *triggering term* in the history $\mathcal{H}_y$ of y, which is itself denoted a *triggered term*. In their adaptation, Seki and Uehara further assumed the triggering terms to be either the subject of an opinion (typically, pronouns) or the object about which the opinion is expressed (typically, concepts in the query). As for the triggered term, it was hypothesized to represent an opinionated expression. The association score $\alpha(y|x)$ for each trigger pattern $x \rightarrow y$ was computed as a maximum likelihood estimate. Using 5,000 customer reviews from Amazon.com, they automatically identified around 10,000 such patterns. Alternatively, in order to adapt these patterns for each individual query, they proposed to recompute the association score $\alpha(y|x)$ using the top blog posts retrieved for the query. Using either the query-independent of query-dependent form of the trigger model, they estimated the probability of a blog post being opinionated, which was then interpolated with the probability $\mathrm{P}_r(q|p)$ of the post being relevant:

$$\mathit{score}(q,p) = (1-\beta)\log \mathrm{P}_r(q|p) + \frac{\beta}{l_p} \sum_{t \in q} \log \mathrm{P}_\zeta(t|h_{t,p}), \tag{3.21}$$

where β is a interpolation parameter controlling the effect of the language model enhanced by opinionated triggers. Using the TREC 2006 Blog track queries, their query-independent model showed a 22% improvement (MAP 0.240) on top of a topic-relevance baseline (MAP 0.196) combining language modelling and inference networks [150], with a 126% gain over the TREC 2006 median (MAP 0.106). Their query-dependent model improved further and significantly (MAP 0.245).

Relatedly, Gerani et al. [61, 64] proposed a context-dependent opinion search approach by aggregating opinionated evidence surrounding multiple occurrences of the query terms. More precisely, for the ith occurrence of a query term t in a retrieved blog post p, their approach computed an *opinion density* $\mathrm{P}(\mathcal{O}|i,p)$ as the estimated

opinionatedness of the context around this occurrence:

$$\mathrm{P}(\mathcal{O}|i,p)\sum_{j=1}^{l_p}\mathrm{P}(\mathcal{O}|t_j)\,\mathrm{P}(j|i,p), \tag{3.22}$$

where l_p is the length of post p, t_j is the term appearing at the jth position of p, $\mathrm{P}(\mathcal{O}|t_j)$ denotes the probability of this term being opinionated, and $\mathrm{P}(j|i,p)$ estimates the probability that the terms at positions i and j are related, according to a kernel $k(j,i)$:

$$\mathrm{P}(j|i,p) = \frac{k(j,i)}{\sum_{j'=1}^{l_p} k(j',i)}. \tag{3.23}$$

To estimate the opinion density at a given position, six kernels were investigated, namely, Gaussian, Laplace, Triangular, Cosine, Circle, and Rectangular. The overall opinion score of the entire blog post was then estimated using different summary statistics, the maximum opinion density being particularly effective:

$$\mathrm{P}(\mathcal{O}|q,p) = \max_{i\in\mathcal{I}_{q,p}} \mathrm{P}(\mathcal{O}|i,p), \tag{3.24}$$

where $\mathcal{I}_{q,p}$ comprises all positions in p that contain a term from q. This formulation was shown to consistently improve upon the five standard baselines provided in the TREC 2008 Blog track [61], with gains of up to 30% (MAP 0.429) over the TREC median (MAP 0.329). Of the six considered kernels, Laplace was found to be the most effective. On the other hand, averaging opinion densities from multiple positions did not perform as well as using only the maximum density. To provide an effective yet more general aggregation mechanism, they further experimented with the ordered weighted averaging (OWA) operator [224]:

$$\mathrm{P}(\mathcal{O}|q,p) = \sum_{i\in\mathcal{I}^*_{q,p}} w_i\,\mathrm{P}(\mathcal{O}|i,p), \tag{3.25}$$

where $\mathcal{I}^*_{q,p}$ once again comprises all positions in p that contain a term from q, but this time sorted in decreasing order of $\mathrm{P}(\mathcal{O}|i,p)$. The weight vector $\vec{w}$, which controls the behavior of the OWA aggregation, was estimated empirically from training data. Their results using the OWA operator significantly outperformed the maximum opinion density, with gains of up to 33% (MAP 0.439) over the TREC median [64].

Another lexicon-based approach was proposed by Vechtomova [214]. In her approach, an opinionated lexicon $\mathcal{L}$ was manually derived from several linguistic resources, including Levin's verb classes denoting psychological state, desire, and judgment (e.g., impress, need, criticise) [118], selected FrameNet's lexical units (e.g., fuss, puzzle, trouble) [17], Ballmer and Brennenstuhl's speech act verbs (e.g., blow up, burst out laughing) [18], Hatzivassiloglou and McKeown's subjective adjectives (e.g., amusing, unreliable) [78], and Wilson's large compilation of subjective terms [222]. In total, the constructed lexicon comprised 10,447 terms. Similarly to Amati et al. [8], each term in the lexicon was further scored based on the KL divergence between its distribution in the set of blog posts assessed as opinionated and the set of all other (non-opinionated) assessed blog posts for the TREC 2006 and 2007 Blog tracks. In order to re-rank the blog posts retrieved by a first-stage topic-relevance baseline, a modified implementation of BM25 [185] was used. In particular, besides taking into account the frequency of the query terms $t \in q$ in each blog post p, this modified version also considered the normalised KL scores of the opinionated terms $t' \in \mathcal{L}$ that co-occurred with t within a window of 30 words. In practice, this was achieved by replacing the standard term frequency component of BM25 with a pseudo-term frequency $\hat{tf}_{t,p}$:

$$\hat{tf}_{t,p} \propto \sum_{i \in \mathcal{I}_{t,p}} \sum_{t' \in \mathcal{L}} \frac{w_{t'}}{\sqrt{dist(t_i, t')}}, \tag{3.26}$$

where $\mathcal{I}_{t,p}$ comprises all positions in p where the term t occurs, $t' \in \mathcal{L}$ is a subjective term from the lexicon $\mathcal{L}$, $w_{t'}$ is given by Equation (3.9), and $dist(t_i, t')$ measures the distance between the ith occurrence of t and the subject term t'. As a result, the proposed approach addressed the requirement that relevant blog posts should contain an expressed opinion toward the query topic. An evaluation showed that the approach performs comparably to the best-performing approaches in the TREC 2007 and 2008 Blog tracks (MAPs 0.429 and 0.423, respectively).

3.2.1.3 Single-Stage Approaches

Differently from the aforementioned two-stage approaches, there have also been attempts to perform opinion search as a single-stage process.

For instance, Zhang and Ye [232] proposed to unify the estimations of topical relevance and opinionatedness into a single generative model for opinion search in blog posts. In their unified model, the relevance and opinionated scores of a post p given the queyr q ($score_r(p|q)$ and $score_o(p|q)$, respectively) were coupled according to:

$$\mathrm{P}(p|q,\mathcal{O}) \propto score_r(p|q)\, score_o(p|q), \tag{3.27}$$

where $score_r(p|q)$ was estimated with BM25 [185], and $score_o(p|q)$ was estimated as the maximum likelihood of a query term co-occurring with a subjective term from the lexicon $\mathcal{L}$, built from SentiWordNet [55]:

$$score_o(p|q) \propto \sum_{t \in q} \sum_{t' \in \mathcal{L}} \frac{pf_{\langle t,t' \rangle, \Upsilon_p}}{\omega\, tf_{t,p}}, \tag{3.28}$$

where $t \in q$ is a query term, $t' \in \mathcal{L}$ is a subjective lexicon term, $pf_{\langle t,t' \rangle, \Upsilon_p}$ denotes the frequency of the pair $\langle t, t' \rangle$ in the set Υ_p of windows of size ω in the post p, and $tf_{t,p}$ is the standard frequency of term t in this post. As a baseline in their investigation, they employed a typical two-stage approach, in which the relevance score and opinionatedness probabilities for each blog post were linearly combined. Their results using the TREC 2007 Blog track queries showed consistent improvements compared to the linear combination baseline under various settings, with gains of up to 28% (MAP 0.337) over a relevance-only baseline (MAP 0.263) and 40% over the TREC 2007 median (MAP 0.242) [134].

In a similar vein, Huang and Croft [89] proposed an alternative single-stage approach to opinion search. However, instead of modelling the document generation process, they proposed to model the user's information need by extending relevance models [113] for both relevance and opinion-oriented query expansion, according to:

$$\begin{aligned} score(q,p) = \alpha \sum_{t \in q} \mathrm{P}(t|\theta_q) \log \mathrm{P}(t|\theta_p) \\ + \beta \sum_{t \in \mathcal{L}_i} \mathrm{P}(t|\theta_{\mathcal{L}_i}) \log \mathrm{P}(t|\theta_p) \\ + \gamma \sum_{t \in \mathcal{L}_q} \mathrm{P}(t|\theta_{\mathcal{L}_q}) \log \mathrm{P}(t|\theta_p), \end{aligned} \tag{3.29}$$

where the first component (parametrized by α) estimates the relevance of the post p to the query q by computing the KL divergence between their language models, θ_p and θ_q, respectively. The other two components, parametrised by β and $\gamma = (1 - \alpha - \beta)$, deploy a query-independent and a query-dependent *sentiment expansion* technique, respectively. 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"), opinionated corpora, such as General Inquirer [201] and OpinionFinder's subjective lexicon [221], and relevance data from the TREC 2006 Blog track. The latter enabled a simple data-driven scheme for selecting effective opinionated terms. In particular, each candidate term t in this multi-source lexicon $\mathcal{L}_i$ was weighted by its average contribution to the opinion search performance of a retrieval system on the set of training queries $\mathcal{Q}$, when t was added to each $q \in \mathcal{Q}$:

$$\mathrm{P}(t|\theta_{\mathcal{L}_i}) \propto \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} (AP(q \cup \{t\}) - AP(q)), \tag{3.30}$$

where $AP(\bullet)$ denotes the average precision attained by a given query formulation. Additionally, a query-dependent sentiment expansion was also proposed to automatically extract expansion terms from a set $\mathcal{O}_q$ of known (or assumed) opinionated blog posts for each query q, in a typical (pseudo-)relevance feedback fashion:

$$\mathrm{P}(t|\theta_{\mathcal{L}_q}) \propto \sum_{p \in \mathcal{O}_q} \mathrm{P}(p)\,\mathrm{P}(t|p) \prod_{t' \in q} \mathrm{P}(t'|p,t), \tag{3.31}$$

where $\mathrm{P}(p)$ was assumed uniform, $\mathrm{P}(t|p)$ was estimated using Dirichlet smoothing [231], and $\mathrm{P}(t'|p,t)$ was estimated as the maximum likelihood of observing the term $t' \in q$ in the post p of length l_p, provided that the lexicon term $t \in \mathcal{L}_q$ had also been observed:

$$\mathrm{P}(t'|p,t) = \begin{cases} tf_{t',p}/l_p & \text{if } tf_{t,p} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.32}$$

Results using a combination of queries from the TREC 2007 and 2008 Blog tracks showed improvements of up to 19% (MAP 0.315) compared

to a relevance-only baseline (MAP 0.265) using the standard relevance model [113] with Dirichlet smoothing [231].

3.2.2 Classification Approaches

Classification-based opinion search approaches build a classifier using training data from sources known to contain either subjective or objective content. The trained classifier is then used to estimate the opinionatedness of blog posts in the target collection. This approach was investigated by Zhang et al. [238], who trained an SVM classifier using data from multiple sources. Subjective training data was obtained from two consumer review Web sites: RateItAll[3] and Epinions.com.[4] Objective training data was obtained from Wikipedia. Classification was performed at the sentence level using both unigram and bigram features, pruned using Pearson's chi-squared test. Moreover, two sentence classification modes were investigated: online and offline. In the online mode, training data was obtained and a classifier was built for each individual test query in a lazy fashion. In the offline mode, training data was obtained and a single classifier was built *a priori* for a set of training queries. In particular, a blog post p was considered opinionated if and only if it contained at least one sentence $\upsilon \in \hat{\Upsilon}_p$ classified as subjective and co-located within a window of five sentences with the query terms or their synonyms. Two approaches were proposed to estimate the opinionatedness of p. The first of these simply counted the total number of sentences in p that met the aforementioned criteria:

$$score_o(q,p) = |\hat{\Upsilon}_p|. \tag{3.33}$$

The second approach considered instead the total confidence of the classification of all sentences $\upsilon \in \hat{\Upsilon}_p$ by the SVM classifier:

$$score_o(q,p) = \sum_{\upsilon \in \hat{\Upsilon}_p} score_{SVM}(\upsilon). \tag{3.34}$$

Using the TREC 2006 Blog track queries, the first approach was found to perform better. Likewise, the offline classification mode performed

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

better than the online mode, with gains of up to 38% (MAP 0.253) over a relevance-only baseline deploying BM25 with phrasal search (MAP 0.183), and 139% over the TREC 2006 median (MAP 0.106). This approach was later refined based on a second-tier classification on top of the initially classified sentences [234]. In particular, the output of the sentence-level SVM classifier was used to produce several post-level features for a decision tree classifier, which classified the entire blog posts. As a result, they achieved further improvements.

Another effective classification-based approach was proposed by He et al. [80]. Their approach used OpinionFinder [222], a subjectivity analysis system aimed to support NLP applications, by providing 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, tokenization, 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 was trained on sentences automatically generated from a large corpus of unannotated data by two rule-based classifiers. In their approach, He et al. computed an opinion score for each retrieved post p based on the proportion of sentences in p that OpinionFinder classified as subjective, and also on the overall confidence of such classification [80]:

$$score_o(q,p) \propto \frac{|\hat{\Upsilon}_p|}{|\Upsilon_p|} \sum_{\upsilon \in \hat{\Upsilon}_p} score_{OF}(\upsilon), \tag{3.35}$$

where $\hat{\Upsilon}_p \subseteq \Upsilon_p$ denotes the subset of all sentences Υ_p in the post p that were classified as subjective. To obtain the final scoring function, this opinion score was then integrated with the initial relevance score:

$$score(q,p) = \frac{-k}{\log \frac{score_o(q,p)}{\sum_{p' \in \mathcal{D}_q} score_o(q,p')}} + score_r(q,p), \tag{3.36}$$

where k is a free parameter, and the previously computed opinion score for the post p is normalized by the sum of the scores of all posts $p' \in \mathcal{D}_q$ retrieved for the query q. Using the query set from the TREC 2007

Blog track [134] for testing, they showed improvements of up to 17% (MAP 0.330) on top of a relevance-only baseline (MAP 0.282) using the DFR PL2F model [76] and 37% compared to the TREC median (MAP 0.242). While these results were obtained by parsing the top 1,200 retrieved blog posts with OpinionFinder, they also showed that significant improvements on top of the strong PL2F baseline could be obtained by parsing as few as 60 posts, substantially contributing to the overall efficiency of their proposed approach.

Zhang et al. [239] proposed a simple classification-based opinion search approach. In their approach, blog posts were represented as vectors over opinionated words, drawn from a manually-annotated subjective lexicon [221]. An SVM classifier was then trained on a pool of blog posts induced from the TREC 2006 Blog track relevance assessments in two different modes: query-independent and query-dependent. In the query-independent mode, posts were sampled from the relevance assessments of all 50 TREC 2006 queries [171]. In the query-dependent mode, a different pool was sampled from the assessments of each of the 50 queries. Cross-validation results on the TREC 2006 queries showed that the query-dependent mode was significantly more accurate than the query-independent mode. Moreover, the accuracy of the trained classifier on the TREC 2007 queries dropped significantly compared to its cross-validation accuracy on the TREC 2006 queries, further suggesting that topically-related posts (i.e., posts retrieved for queries of similar topics) also shared similar sentiment words.

Gerani et al. [62] investigated machine learning approaches for the different steps of an opinion search system. Using all queries from the TREC 2006, 2007, and 2008 Blog tracks under a 10-fold cross validation, they analyzed the performance of an opinion search system at feature selection, blog post classification, and relevance-opinion score aggregation. For feature selection, several metrics were evaluated to select the most effective unigram features, including relative frequency, likelihood ratio, mutual information, and chi-squared statistics, with the latter two yielding the best performances. For blog post classification, they experimented with a simple average of the scores of all terms in the post, as well as with an SVM classifier using the frequency of the top 2% scored terms as features. The latter approach was found

to perform better, attaining an overall high recall but low precision, suggesting that higher-order n-gram features could have been beneficial. Finally, to aggregate the obtained opinion score with the relevance score from a baseline system, they investigated unsupervised aggregation methods, such as BordaFuse [158]:

$$rank(q,p) = rank_r(q,p) + rank_o(p,q), \tag{3.37}$$

where $rank_r(q,p)$ and $rank_o(p,q)$ denote the rank position of the post p according to relevance and opinion scoring mechanisms, respectively. Besides unsupervised methods, they investigated a supervised aggregation method using structural SVM to maximize MAP [228]. In this investigation, they found that BordaFuse could improve the relevance-only baseline (MAP 0.354) by up to 11% (MAP 0.392), while the supervised approach using SVM improved up to 50% (MAP 0.531).

Liu et al. [125, 119] proposed to leverage several sentiment features for opinion search, building a lexicon $\mathcal{L}$ with subjective n-grams and part-of-speech tags. Unigram sentiment features were semi-automatically derived from an initial set of seed sentiment words, expanded to include highly co-occurring adjectives in a collection of reviews for movies [179], general products [88], and hotels. After inspection by a native English speaker, the 50 most frequent positive and negative terms were kept as additional sentiment features. Trigram features were then computed based on the co-occurrence of terms or their part-of-speech with sentiment terms. For blog post classification, these sentiment features were first aggregated at the sentence level, by accumulating the mutual information between each adjective t in a sentence v and all the identified positive and negative sentiment features t' in the lexicon $\mathcal{L}$, according to:

$$score_o(q,v) = \sum_{t \in v} \frac{1}{|\mathcal{L}|} \sum_{t' \in \mathcal{L}} pf_{\langle t,t' \rangle, \Upsilon_{\mathcal{O}}}, \tag{3.38}$$

where $pf_{\langle t,t' \rangle, \Upsilon_{\mathcal{O}}}$ denotes the frequency of the pair $\langle t,t' \rangle$ among the sentences $\Upsilon_{\mathcal{O}}$ identified from opinionated training data. Likewise, post-level features were then generated by aggregating the sentence scores:

$$score_o(q,p) = \Psi_{v \in \Upsilon_p} score_o(q,v), \tag{3.39}$$

where Υ_p comprises all sentences v in the post p, and Ψ is a summary statistic, such as the mean positive or negative sentence score, as well as the mean difference and ratio between each sentence's positive and negative scores. Two classification modes were considered with a multinomial logistic regression classifier: binary classification, to distinguish between non-opinionated and opinionated posts, and multi-class classification, to distinguish between non-opinionated, positively opinionated, negatively opinionated, and mixedly opinionated. Using the TREC 2006 Blog track queries under a 5-fold cross validation, they showed that the trigram features were beneficial for the multi-class, but not for the binary classification mode. Indeed, the binary classification mode was shown to generally outperform the multi-class classification [125]. By integrating the classification posterior probability as a prior for ranking blog posts, they attained a modest improvement (0.6%, MAP 0.384) on top of the best-performing standard baseline from the TREC 2008 Blog track (MAP 0.382) [173].

3.2.3 Polarity Approaches

When presenting opinionated blog post search results, it can be advantageous for the users to be shown a selection of the differing opinions surrounding the query topic. The most natural way to illustrate the opinions is by separating the negative from the positive ones [179, 212], so that the *polarity* of opinions can be clearly ascertained. Most of the approaches described in the previous sections focused on identifying opinionated blog posts, regardless of the polarity of the opinion expressed in these posts. While some of these approaches are also applicable for polarity search, in the following, we describe approaches that explicitly tackled the problem of polarity orientation.

For instance, Chenlo and Losada [38] proposed to estimate the polarity of blog posts using sentence-level evidence. Their intuition was that blog posts could contain mixed opinions following some predictable flow (e.g., with a definitive opinion at the beginning or at the end of the post). Accordingly, they estimated the polarity of each sentence v in a blog post as a linear combination between the estimated sentence relevance $score_r(q, v)$, as given by BM25 [185], and the fraction of positive

(resp., negative) terms in the sentence, as tagged in OpinionFinder's subjective lexicon $\mathcal{L}_+$ (resp., $\mathcal{L}_-$) [222]:

$$score_+(q,\upsilon) = \beta\, score_r(q,\upsilon) + (1-\beta)\frac{\sum_{t\in\mathcal{L}_+} tf_{t,\upsilon}}{l_\upsilon}, \tag{3.40}$$

where l_υ denotes the length (in tokens) of the sentence υ and β is a free parameter. The polarity of each blog post was then obtained by aggregating sentence polarity scores, according to:

$$score(q,p) = \gamma\, score_r(q,p) + (1-\gamma)\Psi_{\upsilon\in\Upsilon_p}\, score_+(q,\upsilon), \tag{3.41}$$

where $score_r(q,p)$ is the relevance score of post p with respect to the query q, as given by a topic-relevance baseline, Ψ is a summary statistic aggregating the polarity scores of the sentences $\upsilon \in \Upsilon_p$, and γ is a parameter controlling the balance between relevance and polarity. Different summary statistics were considered, including the average score of all sentences, or just the best, first, or last k sentences. Using the TREC 2008 Blog track queries, they showed that their proposed approach could consistently outperform all five standard relevance-only baselines, as well as the opinion search approach of He et al. [80], which also used OpinionFinder, but without focusing on polarity. Moreover, their attained performance was comparable to that of the best performing approach at the TREC 2008 Blog track [173].

Jia et al. [91] investigated the impact of negation terms on opinion search performance. In particular, they proposed a set of syntactic rules for identifying the scope of a negation term. Opinionated terms occurring within the scope of a negation had their polarity flipped. As a result, such negation-aware opinionated terms could be used to improve the accuracy of the classification of the polarity of entire sentences or posts. Experiments on the TREC 2006, 2007, and 2008 Blog track queries showed that their proposed approach outperformed several heuristic methods in the literature for negation scope identification.

A different approach for polarity search was introduced by Demartini [48]. In particular, instead of presenting users with results of a single polarity inclination, he proposed to return a ranking covering diverse polarities, so as to provide the searchers with multiple viewpoints about their search topic. In order to promote diversity in the

ranking, he deployed the xQuAD framework for search result diversification [191]. In particular, xQuAD diversifies the search results for a query q in a greedy, iterative fashion. In each iteration, it selects the result p that maximizes the following probabilistic objective:

$$score(q, p, \mathcal{D}') = (1 - \lambda)\,\mathrm{P}(p|q) + \lambda\,\mathrm{P}(p, \bar{\mathcal{D}}'|q), \tag{3.42}$$

where $\mathcal{D}'$ comprises the results selected in the previous iterations of the algorithm, $\mathrm{P}(p|q)$ and $\mathrm{P}(p, \bar{\mathcal{D}}'|q)$ denote the probability of p being relevant and diverse, respectively, and λ controls the trade-off between these two probabilities. The former probability can be estimated by any standard retrieval approach. In order to estimate $\mathrm{P}(p, \bar{\mathcal{D}}'|q)$, xQuAD assumes that the query q may represent not one, but multiple aspects of the user's information need, which results in the following definition:

$$\mathrm{P}(p, \bar{\mathcal{D}}'|q) = \sum_{a \in \mathcal{A}} \Big[\mathrm{P}(a|q)\,\mathrm{P}(p|q, a)\,\mathrm{P}(\bar{\mathcal{D}}'|q, a) \Big], \tag{3.43}$$

where $\mathcal{A}$ denote the possible aspects underlying the query q. In the context of polarity search, query aspects were defined as "positive," "negative," or "mixed." The probability $\mathrm{P}(a|q)$ denotes the likelihood of each aspect, and was assumed uniform. To estimate the probability $\mathrm{P}(p|q, a)$ that a blog post p covers the aspect a, as well as the probability $\mathrm{P}(\bar{\mathcal{D}}'|q, a)$ that none of the already selected posts in $\mathcal{D}'$ covers this aspect, they deployed a sentence-level SVM classifier trained using the relevance assessments from the polarity task of the TREC 2008 Blog track [173]. In particular, $\mathrm{P}(p|q, a)$ was estimated as:

$$\mathrm{P}(p|q, a) \propto \frac{\sum_{\upsilon \in \Upsilon_p} \mathrm{SVM}_a(\upsilon)}{|\Upsilon_p|}, \tag{3.44}$$

where Υ_p comprises all sentences in the post p that contain the query terms, and $\mathrm{SVM}_a(\upsilon)$ denotes the confidence of the classification of the sentence $\upsilon \in \Upsilon_p$ as being of polarity a. Qualitative results showed the promise of the proposed approach, particularly for controversial queries such as "death penalty" or "termination of pregnancy."

3.3 Summary

This section has surveyed a wide number of approaches related to the search of blog posts, ranging from indexing issues such as text cleaning

(boilerplate removal) to the identification of the opinionated nature of the blog posts, or even their polarity orientation toward a given query topic. Such queries typically aim to identify the context around the occurrences of a named entity (context queries) — as highlighted in Section 2.3. Of course, blog posts are just one element of the semi-structured nature of the blogosphere, and context queries are one form of information need on the blogosphere. In the next section, we review approaches that are able to identify whole blogs that are of interest to a user issuing a concept query.

4

Blog Search

While blogs may be written in either a personal or a more official 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 they would like to read on a regular basis (for instance, the authors of this survey are interested in information retrieval blogs). Indeed, as discussed in Section 2, a study of the query log of a blog search engine revealed *concept queries*, where users seemed to be looking for blogs to subscribe to [156]. In this section, we survey approaches that tackle several aspects of blog search. In particular, Section 4.1 describes approaches devoted to identifying relevant blogs, represented as either pseudo-documents or aggregates of blog posts. Section 4.2 discusses existing approaches for (pseudo-)relevance feedback for blog search. Section 4.3 describes approaches that exploit the dynamic nature of the blogosphere, by trying to identify blogs that recurrently write about a topic of interest. Section 4.4 describes several attempts to measure the prior relevance of a blog post regardless of any query, including the identification of authoritative or influential bloggers. Finally, Section 4.5 discusses blog search in light of multiple facets of interest, so as to facilitate the access to ranked blogs fulfilling different criteria.

4.1 Topical Relevance

In Section 3.1.3, we described several approaches aimed at estimating the relevance of a blog post with respect to the user's query. Estimating the relevance of an entire blog to a query, on the other hand, is intrinsically more challenging, as each blog comprises multiple posts, each of which with a different degree of relevance itself. This granularity issue is also what essentially differentiates the most effective approaches in the literature for estimating the topical relevance of blogs. In the remainder of this section, we describe these approaches, organized according to the search tasks that inspired them.

4.1.1 Resource Selection Approaches

The view of blogs as collections of posts directly links the blog search task to the resource selection problem in distributed information retrieval [32], in which the goal is to select the collections (or resources) more likely to contain documents relevant to a given query. Elsas et al. [52] were among the first to explore the link between the two tasks by proposing different models for representing blogs, so as to make an explicit distinction between using the large documents (blogs) versus the small documents (blog posts) views of blog search. The first representation, called the *Large Document* (LD) model, treated a blog b as a concatenation of all its posts $p \in b$. This combined representation was used to rank blogs by their posterior probability given the query q:

$$\mathrm{P}_{LD}(b|q) \propto \mathrm{P}(b)\,\mathrm{P}(q|b), \tag{4.1}$$

where the prior $\mathrm{P}(b)$ was set proportionally to the number of posts n_b in the blog b, i.e., $\mathrm{P}(b) \propto \log(n_b)$, and the query likelihood $\mathrm{P}(q|b)$ was estimated with Dirichlet smoothing [231] and a full dependence model, in order to account for term dependencies [151]. An additional enhancement considered not only the HTML content of blog posts, but also their different fields (e.g., blog title, post title) from the blog's syndicated XML feeds [13]. In the second representation, known as the *Small Document* (SD) model, blogs were seen as a collection of blog posts, as in a typical resource selection problem. Accordingly, they adapted 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 b was scored based on a combination of the individual scores of its posts:

$$\mathrm{P}_{SD}(b|q) \propto \mathrm{P}(b) \sum_{p \in b} \mathrm{P}(p|b)\, \mathrm{P}(q|p), \tag{4.2}$$

where the prior $\mathrm{P}(b)$ was computed as before, and the query likelihood $\mathrm{P}(q|p)$ was estimated using Jelinek–Mercer smoothing [231]. Again, individual posts were 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 $\mathrm{P}(p|b)$ was used to infer how well the post represented the language model of the whole blog with respect to the query terms [52]:

$$\mathrm{P}(p|b) = \frac{\phi(p,b)}{\sum_{p' \in b} \phi(p',b)}, \tag{4.3}$$

where $\phi(p,b)$ is a measure of the similarity between the post p and the blog b. For instance, using a geometric mean of term generation probabilities, $\phi(p,b)$ was defined as:

$$\phi(p,b) = \prod_{t \in p} \mathrm{P}(t|\theta_b)^{\mathrm{P}(t|\theta_p)}, \tag{4.4}$$

where the post language model θ_p was estimated through a maximum likelihood estimate, and the blog model θ_b was estimated as:

$$\mathrm{P}(t|\theta_b) = \frac{1}{n_b} \sum_{p \in b} \mathrm{P}(t|\theta_p), \tag{4.5}$$

where n_b is the number of posts in b. Using the query set of the blog distillation task of the TREC 2007 Blog track [134], they showed that, with a uniform prior $\mathrm{P}(b)$, SD and LD perform similarly (MAP 0.290). However, when this prior was set proportionally to the number of posts in a blog, SD (MAP 0.315) markedly outperformed LD (MAP 0.188), with gains of up to 55% over the TREC 2007 median (MAP 0.203) [134].

Seo and Croft [195] also approached blog search as a resource selection problem, by seeing blogs as collections of blog posts. Similarly to the approach of Elsas et al. [52], they also considered different representations of blogs, namely, a *Global Representation* (GR), and an

alternative representation, called *Pseudo-Cluster Selection* (PCS). The GR model treated blogs as a concatenation of all their posts, as in the LD model of Elsas et al. In particular, the likelihood $\mathrm{P}_{GR}(q|b)$ of a blog b generating the query q was estimated as:

$$\mathrm{P}_{GR}(q|b) = \prod_{t \in q} \mathrm{P}(t|\theta_b)^{tf_{t,b}}, \tag{4.6}$$

where $\mathrm{P}(t|\theta_b)$ deployed a maximum likelihood estimate with Dirichlet smoothing [231]. Their second representation, PCS, was analogous to the SD model of Elsas et al., but was based on a different principle. In PCS, a blog b was seen as a query-dependent cluster containing only highly ranked blog posts for a given query q:

$$\mathrm{P}_{PCS}(q|b) \propto \left[\prod_{j=1}^{k} \prod_{t \in q} \mathrm{P}(t|\theta_{p_j}) \right]^{\frac{1}{k}}, \tag{4.7}$$

where the parameter k denoted the ideal number of highly ranked posts to be considered from each blog and $\mathrm{P}(t|\theta_{p_j})$ also deployed a Dirichlet-smoothed maximum likelihood estimate. Considered separately, GR was shown to outperform PCS on the query set of the blog distillation task of the TREC 2007 Blog track (MAP 0.345 vs. 0.317), with improvements of 70% over the TREC median (MAP 0.203) [134]. Additionally, a strategy that combined the two models was shown to outperform both individually, hence demonstrating their complementary characteristics [195]. Indeed, the global nature of GR helped uncover the prevailing topics covered by a blog instead of a multitude of relatively less important topics, whereas the selection strategy employed by PCS mitigated 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 penalizing topical diversity while reducing the overhead of having a second index, they created a query-independent version of PCS, by randomly sampling posts from each blog. Furthermore, in order to focus on the temporal aspect of blogs, where those with more recent posts on a given topic are more likely to be relevant to the topic, this sampling

was biased toward recently added posts. This alternative version was shown to perform comparably to the GR model [195].

A third approach to blog search built upon the two previously described approaches. In particular, Lee et al. [115] adapted the SD model of Elsas et al. [52] — renamed the *Global Evidence Model* (GEM) — and the PCS model of Seo and Croft [195] — renamed the *Local Evidence Model* (LEM) — in the context of the risk minimisation framework of Lafferty and Zhai [111]. In practice, both GEM and LEM were implemented identically, except that GEM considered every post p in a given blog b, whereas LEM only considered the top retrieved posts:

$$score_{GEM}(q,b) = \frac{1}{n_b} \sum_{p \in b} score(q,p), \tag{4.8}$$

$$score_{LEM}(q,b) = \frac{1}{n_{\hat{b}}} \sum_{p \in \hat{b}} score(q,p), \tag{4.9}$$

where $\hat{b}$ denotes the top $n_{\hat{b}}$ retrieved posts for the query q that belong to blog b. Besides being based on a different framework, their approach directly addressed two weaknesses of both SD [52] and PCS [195] when considered individually. Firstly, to overcome the problem of blogs being overly represented by a few long posts, all blog posts were considered equally important to the blog they belong to, which was expressed in their probabilistic framework as a uniform probability of posts being retrieved given their blog. Secondly, to avoid a bias toward prolific blogs, the score of a given blog was computed as the average score of its posts. A linear combination of the scores produced by GEM and LEM achieved the top performance among the participant approaches in the TREC 2008 Blog track, with gains of up to 35% (MAP 0.301) over the TREC 2008 median (MAP 0.224) [173].

An alternative blog search approach inspired by resource selection was proposed by Keikha and Crestani [95, 96]. Their approach leveraged Ordered Weighted Averaging (OWA) operators, a parametrized class of mean aggregation operators [224]. In particular, given a query q, the score of a blog b was estimated as:

$$score(q,b) = \sum_{i=1}^{n_b} w_i \, score(q,p_i), \tag{4.10}$$

where p_i is the ith highest scored post in blog b, from a total of n_b posts, and the weight vector $\vec{w}$ determines the behavior of the aggregation. For instance, depending on its associated weighting vector, an OWA operator can behave more like an OR operator or an AND operator. In order to obtain an effective weight vector, Keikha and Crestani experimented with fuzzy linguistic quantifiers. In practice, the weight w_i of the ith highest scored post p_i was estimated as:

$$w_i = \eta\left(\frac{i}{n_b}\right) - \eta\left(\frac{i-1}{n_b}\right), \tag{4.11}$$

where $\eta(\bullet)$ is a fuzzy linguistic quantifier [229], defined as:

$$\eta(r) = \begin{cases} 0, & \text{if } r < a, \\ \frac{r-a}{b-a}, & \text{if } a \leq r \leq b, \\ 1, & \text{if } r > b, \end{cases} \tag{4.12}$$

where a and b are free parameters, with $a, b, r \in [0, 1]$. With different values for a and b, they manually defined different quantifiers for determining the actual posts that would play a role in the score aggregation, such as "at least half" ($a = 0.0, b = 0.5$), "most" ($a = 0.3, b = 0.8$), and "as many as possible" ($a = 0.5, b = 1.0$). Results using the TREC 2007, 2008, and 2009 Blog track queries showed improvements of up to 53% (MAP 0.312), 8% (MAP 0.242), and 63% (MAP 0.210) over the median performances of the corresponding years (MAPs 0.203, 0.224, and 0.128, respectively) [134, 173, 136]. Their best performing setting revealed that favoring blogs with highly scoring posts was an effective ranking strategy [95]. In order to further improve the original OWA operators, Keikha and Crestani [96] proposed to account for the importance of each retrieved blog post, estimated via a random walk on the term-post graph [45], a bipartite graph with nodes representing the query terms and the top retrieved posts for the query. These extended operators were shown to bring further improvements, particularly when aggregating evidence from a large number of posts from each blog.

4.1.2 Expert Search Approaches

A different class of approaches for blog search explored the similarities between this task and the expert search task [16]. In the 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 [133] were the first to propose tackling blog search as an expert search problem, by adapting their expert search model — the so-called *Voting Model* [132] — 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 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 a blog post p retrieved for a given query q that belongs to the profile of a blog b is considered as a vote for the relevance of that blog to the query. More formally:

$$score_{VM}(q,b) = \Psi_{p\in\hat{b}} f(q,p), \tag{4.13}$$

where Ψ is a summary statistic (e.g., sum, average, max, count) and f is a function of the query q and each post $p \in \hat{b}$, where $\hat{b} = \mathcal{D}_q \cap b$ denotes the subset of the posts in b that are retrieved for q. Different instantiations of Ψ and f result in different voting techniques, which provide different ways of converting a ranking of blog posts into a ranking of blogs. For instance, Macdonald and Ounis showed gains of up to 27% (MAP 0.258) over the TREC 2007 median (MAP 0.203) [134] using their expCombMNZ voting technique [133]:

$$score_{expCombMNZ}(q,b) = n_{\hat{b}} \sum_{p\in\hat{b}} \exp(score(q,p)), \tag{4.14}$$

where $n_{\hat{b}}$ is the number of posts from blog b retrieved for the query q, $\exp(\bullet)$ denotes the exponential function, which favors highly scored blog posts more aggressively, and $score(q,p)$ can be estimated by any standard ranking model. Additionally, they showed that techniques intended to enhance the underlying ranking of blog posts, such as taking into account query term occurrences in close proximity, resulted in an improved blog search effectiveness. Besides enhancing the ranking of blog posts, techniques applied to the ranking of blogs were

shown to bring additional improvements. These included a normalization technique called Norm2 to counterbalance an eventual bias toward prolific bloggers, which are more likely to be retrieved for any query [137]:

$$score_{VM+Norm2}(q,b) = score_{VM}(q,b) \log\left(1 + \frac{c\bar{n}}{n_b}\right), \tag{4.15}$$

where $c > 0$ is a parameter controlling the amount of normalization, $\bar{n}$ is the average number of posts across all blogs, and n_b is the total number of posts in blog b. With these combined techniques, further gains of up to 68% (MAP 0.342) over the TREC median were observed [133].

The connection between blog search and expert search was also explored by Balog et al. [20]. In particular, they adapted two language modelling approaches originally developed for expert search [19] in order to search for relevant blogs. In their *Blogger Model* (BM), the probability of a query q being generated by a given blog b was estimated by representing this blog as a multinomial distribution over terms:

$$\mathrm{P}(q|\theta_b) = \prod_{t \in q} \mathrm{P}(t|\theta_b)^{tf_{t,q}}, \tag{4.16}$$

where $\mathrm{P}(t|\theta_b)$ was computed as a Jelinek–Mercer-smoothed estimate of $\mathrm{P}(t|b)$. The latter probability was estimated as:

$$\mathrm{P}(t|b) = \sum_{p \in b} \mathrm{P}(p|b)\,\mathrm{P}(t|p,b), \tag{4.17}$$

where $\mathrm{P}(t|p,b)$ was computed as a maximum likelihood estimate of $\mathrm{P}(t|p)$, assuming that p and b were conditionally independent. The probability $\mathrm{P}(p|b)$ of choosing a post p given its blog b was assumed uniform. In the *Posting Model* (PM), the query likelihood given a blog b was estimated by combining the estimated probabilities that the individual blog posts in this blog would generate the query:

$$\mathrm{P}(q|b) = \sum_{p \in b} \mathrm{P}(p|b) \prod_{t \in q} \mathrm{P}(t|\theta_p)^{tf_{t,q}}, \tag{4.18}$$

where, similarly to BM, $\mathrm{P}(p|b)$ was assumed uniform and $\mathrm{P}(t|\theta_p)$ was obtained via Jelinek–Mercer smoothing of $\mathrm{P}(t|p)$. Using the TREC 2007 Blog track queries, they showed that BM (MAP 0.327) significantly outperformed PM (MAP 0.232), with gains of 61% over the TREC

median performance (MAP 0.203). Interestingly, they noted that the relative performance of the counterpart expert search models of BM and PM was the other way around [19]. This reinforced the observation that, differently from the expert search task, the blog search task requires bloggers to not only write about the topic of interest, but to do so on a focused and recurrent basis. Finally, the efficiency of BM was later improved by Weerkamp et al. [219]. In particular, they showed that re-ranking a selection of blogs — namely, those with at least one post among the top retrieved posts for the query — was at least as effective as ranking all blogs (i.e., the standard BM), while requiring the examination of as few as 1% of all blog-post associations.

4.2 Relevance Feedback

Similarly to a blog post search scenario, relevance feedback techniques have also been applied in the context of blog search. Typically, they take into account the hierarchical structure of blogs and their posts as feedback items, as well as the availability of higher quality resources outside the blogosphere for obtaining expansion terms. For instance, Elsas et al. [52] investigated several mechanisms for query expansion on the blogosphere, including traditional pseudo-relevance feedback using relevance models [113] and a novel link-based expansion approach. The former employed both the target corpus (blogs and blog posts) as well as Wikipedia articles as feedback items. The latter built upon a Wikipedia ranking for a given query q, considered as two overlapping portions: a *working* portion $\mathcal{D}_k$, comprising the top k retrieved articles, and a *relevant* portion $\mathcal{D}_r$, comprising the top r articles, with $r < k$. Each anchor phrase — the clickable text in a hyperlink — $a \in \Upsilon_{\mathcal{D}_k}$ that referred to articles in the relevant portion $\mathcal{D}_r$ was scored as:

$$score(a) = \sum_{a_i \in \Upsilon_{\mathcal{D}_k}} \left[\mathbf{1}_{\mathcal{D}_r}(target(a_i)) \; (r - rank(target(a_i), \mathcal{D}_r)) \right], \quad (4.19)$$

where a_i denotes an occurrence of the anchor phrase a in the set $\Upsilon_{\mathcal{D}_k}$ of all unique anchors from the working portion $\mathcal{D}_k$, the function $target(a_i)$ returns the target article of the hyperlink anchored by a_i, $\mathbf{1}_{\mathcal{D}_r}(target(a_i))$ is the indicator function, which determines whether the target article belongs to $\mathcal{D}_r$, and $rank(target(a_i), \mathcal{D}_r)$ indicates

the position of the Wikipedia article pointed by a_i in the ranking $\mathcal{D}_r$. From the highest scored anchor phrases, the top 20 were selected to expand the original query. Results using the TREC 2007 queries showed that traditional pseudo-relevance feedback could improve upon a non-expanded baseline using the LD representation (MAP 0.327 vs. 0.290), but not the SD representation (MAP 0.314 vs. 0.315), as described in Section 4.1.1. In turn, the link-based expansion approach using Wikipedia improved upon both the LD and SD representations, with gains of up to 22% (MAP 0.355) and 15% (MAP 0.361), respectively [52].

Lee et al. [116] proposed to address two problems that emerge from the direct application of pseudo-relevance feedback on the blogosphere: the topic bias incurred by expanding terms from highly ranked blog posts, and the topic drift incurred by expanding terms from all posts of each blog. To overcome both problems, they proposed a diversity-oriented query expansion approach. Differently from a traditional application of pseudo-relevance feedback, their approach considered the top m retrieved posts from the top k retrieved blogs as the pseudo-relevance feedback set, defined as $\mathcal{F}_q = \{p_{i,j} \mid i = 1 \cdots k, j = 1 \cdots m\}$. 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 provided a richer vocabulary around the topic of the query, which had the potential to produce a more effective expanded query. At the same time, by focusing on highly scored posts within each blog, topic drift was also handled. An empirical evaluation using a combination of the GEM and LEM models described in Section 4.1.1 as a baseline showed improvements of up to 9% (MAP 0.423) on the TREC 2007 Blog track queries [134] and 8% (MAP 0.325) on the TREC 2008 queries [173] (baseline MAPs 0.393 and 0.301, respectively).

4.3 Temporal Relevance

One of the distinctive characteristics of blog search compared to other search tasks is its temporal nature. Indeed, blog searchers are typically interested in blogs with a *recurrent interest* on a particular topic, so that they can subscribe to these blogs' syndicated feed and follow their

posts on a regular basis [156]. With this implicit requirement in mind, Macdonald and Ounis [133] proposed an approach to favor blogs that recurrently mention the topic of the query. In particular, they postulated that blogs with a recurrent interest on the topic of the query would have relevant posts spread across a long timespan. To quantify this notion for each blog b, they computed the monthly fraction of posts from the blog among the top retrieved posts for the query:

$$score_{Dates}(q,b) = \sum_{i=1}^{3} \frac{1 + n_{\hat{b}_i}}{1 + n_{b_i}}, \tag{4.20}$$

where n_{b_i} denotes the number of posts from the blog b that were produced in the ith time interval, and $n_{\hat{b}_i}$ is the number of such posts that were among the top retrieved posts for the query q. The number of monthly intervals was set to 3, as the underlying corpus spanned 11 weeks. This recurrence score was linearly combined with the relevance score produced by the expCombMNZ voting technique within their Voting Model (MAP 0.258), described in Section 4.1.2. Experiments using the TREC 2007 Blog track queries showed an improvement of 15% (MAP 0.298) by accounting for the recurrent interest of the retrieved blogs with respect to the topic of the query.

Similarly, Keikha et al. [97] proposed a framework to measure the stability of a blog's relevance over time. In particular, they estimated the relevance stability of a blog b as the likelihood of observing relevant content from this blog across fixed-size time windows v:

$$score_{Stability}(q,b) = \sum_{v} \mathrm{P}(v|q)\,\mathrm{P}(b|q,v). \tag{4.21}$$

This estimation involved two main components: the window importance $\mathrm{P}(v|q)$ and the window-based query likelihood $\mathrm{P}(b|q,v)$. For the window importance component, they experimented with a uniform estimate, as well as with an estimate proportional to the total score of the blog posts produced during the window timespan:

$$\mathrm{P}(v|q) \propto \sum_{p \in (\cup_v b_v)} \mathrm{P}(p|q), \tag{4.22}$$

where $\cup_v b_v$ denotes the union of the posts produced within the window v from all blogs, and $\mathrm{P}(p|q)$ was estimated using Dirichlet

smoothing [231]. To determine the window size, they experimented with both query-independent and query-dependent approaches. In the query-independent approach, the target window size was learned from training data. For the query-dependent approach, the window size was defined as the average distance between two consecutive top retrieved posts for a given query. Finally, the window-based query likelihood $\mathrm{P}(b|q,\upsilon)$ was estimated proportionally to the estimated relevance of the highest scored post $p \in b_\upsilon$ within the window υ:

$$\mathrm{P}(b|q,\upsilon) \propto \max_{p \in (\cup_\upsilon b_\upsilon)} \mathrm{P}(p|q), \tag{4.23}$$

where $\mathrm{P}(p|q)$ was once again estimated using Dirichlet smoothing [231]. Using either approach, the final stability score was linearly combined with the relevance score produced by the SD model of Elsas et al. [52], which served as a baseline in their investigation (MAP 0.205). Another baseline used the recurrence approach of Macdonald and Ounis [133] in combination with SD (MAP 0.210). Cross-validation results using the TREC 2009 Blog track queries [136] showed that the proposed stability approach significantly outperformed both baselines (MAP 0.236), with gains of up to 15% and 12%, respectively [97].

In addition to improving relevance estimations, Keikha et al. [98] also proposed to leverage temporal information to improve query expansion on the blogosphere. Their approach was based on the intuition that a query term might not be a good representation of the user's information need at all times. Accordingly, they proposed to represent both the query and its top retrieved blogs in a temporal space. Using these temporal representations, they proposed different methods for aggregating the temporal relevance scores of a blog b given a query q. Among the proposed methods, the euclidean distance between the blog and query vectors was the most effective:

$$score_{Temporal}(b|q) = \sqrt{\sum_i score(q,b,i)^2}, \tag{4.24}$$

where $score(q,b,i)$ denoted the estimated relevance score at the ith daily interval, computed as the cosine between the vector representations

of b and q at time i, denoted $\vec{b_i}$ and $\vec{q_i}$, according to:

$$score(q, b, i) = \frac{\vec{q_i} \cdot \vec{b_i}}{||\vec{q_i}|| \, ||\vec{b_i}||}, \tag{4.25}$$

where $\vec{q_i}$ comprised the top retrieved blog posts for q at time i, and $\vec{b_i}$ comprised all posts in b from the same time period. Finally, the temporal score of a blog was linearly combined with its relevance score, as produced by the Blogger Model of Balog et al. [20]. Besides BM (MAP 0.277), their approach was also compared to the query expansion approaches proposed by Elsas et al. (MAP 0.289) [52] and Lee et al. (MAP 0.289) [116], as described in Section 4.2. Cross-validation results using the TREC 2009 Blog track queries [136] showed that the proposed approach significantly outperformed all baselines, with gains of up to 12.5% (MAP 0.312) over the non-expanded BM baseline [98].

4.4 Prior Relevance

Up to now, we have discussed blog search approaches that treated relevance as a query-dependent quantity. However, relevance as perceived by a blog searcher may span other criteria beyond the topical one. For instance, blog searchers may be more inclined to subscribe to content produced by authoritative or influential bloggers, regardless of any particular topic. Identifying such bloggers (or their blogs) is important, as they are the ones who affect others' decisions, from buying a new product to discussing social and societal issues [3, 66]. At the same time, besides leading to increased readership and linking, authority and influence contribute to improving a blogger's status on the blogosphere [2].

Traditional approaches to infer the importance of a Web page rely on the hyperlink structure of the Web. However, blogs are very sparsely linked, making the blogosphere graph inadequate for algorithms based upon a random surfer model [108], such as PageRank [175] and HITS [101]. While the authority of a blog may improve over time as it receives more links, its influence tends to diminish, as the blogosphere as a whole gets sparser with the addition of thousands of new sparsely-linked blog posts [4]. To combat the sparsity of the link

structure underlying the blogosphere, Adar et al. [2] proposed to infer implicit links between blogs. Several features were proposed for the automatic prediction of implicit links between blogs b_i and b_j using SVM. These included the number of links to blog and non-blog pages shared by b_i and b_j, their textual similarity, and the likelihood that b_j relays links first posted by b_i. An evaluation using daily differential crawls of BlogPulse from May 2003 showed a cross-validation accuracy of 91% at distinguishing between linked and unlinked blogs. Notably, relying solely on whether two blogs shared a common link yielded an accuracy of 88%. However, predicting the link direction turned out to be less effective, with an accuracy of 57%. To investigate the usefulness of this implicit link structure for blog search, they further compared the ranking of blogs produced by the PageRank algorithm [175] on the explicit (hyperlink) structure of the crawl to the ranking produced on its implicit structure using the same algorithm, renamed iRank. A qualitative analysis showed that while PageRank accurately identified authoritative blogs (i.e., those that create information), iRank was more effective at identifying influential blogs (i.e., those that spread information).

A related approach to overcome the sparse link structure of the blogosphere was proposed by Fujimura et al. [58], who introduced the EigenRumor algorithm to rank blogs. The algorithm operated on a bipartite graph linking blogs and blog posts. A link g_{ij} between a blog b_i and one of its posts p_j was denoted a provisioning link, while a link e_{ij} from b_i to an outside post p_j was called an evaluation link, emphasizing the blogger's support to the contents of the linked posts. Both links were weighted inversely proportionally to their age at time τ, to model the decrease of interest in older posts, according to:

$$g_{ij}^{(\tau)} = \frac{\rho^{\tau - time(g_{ij})}}{\sqrt{\sum_{k=1}^{n_{b_i}} \rho^{\tau - time(g_{ik})}}} \quad \text{and}$$

$$e_{ij}^{(\tau)} = \frac{\gamma^{\tau - time(e_{ij})}}{\sqrt{\sum_{k=1}^{n_{b_i}} \gamma^{\tau - time(e_{ik})}}}, \tag{4.26}$$

where n_{b_i} denotes the number of posts in blog b_i, ρ and γ are damping factors determined empirically and $time(\bullet)$ denotes the creation time

of a given link. From the produced bipartite graph, their proposed algorithm iteratively computed *hub* (evaluation) and *authority* (provisioning) scores for blogs — similarly to the HITS algorithm [101] — and a *reputation* score for blog posts. The authority score of a blog was computed as a function of the reputation of its provided blog posts, while its hub score was computed as a function of the reputation of the posts evaluated by the blogger. More precisely:

$$\vec{a} = \vec{g} \cdot \vec{r}, \tag{4.27}$$

$$\vec{h} = \vec{e} \cdot \vec{r}, \tag{4.28}$$

where $\vec{a}$ and $\vec{h}$ denoted the resulting authority and hub vectors, respectively, while $\vec{r}$ denoted the reputation of each post as a function of the authority score of its provider and the hub score of its evaluator:

$$\vec{r} = \alpha(\vec{g}^T \cdot \vec{a}) + (1 - \alpha)(\vec{e}^T \cdot \vec{h}), \tag{4.29}$$

where $\alpha \in [0, 1]$ is a parameter that controls the balance between the authority and hub scores, which could be adjusted depending on the target application. To evaluate the proposed algorithm, they used a 3.5-week crawl of the Japanese blogosphere, with over 9 m posts from 300 k blogs, of which only 1.15% were referred to by other blogs, and hence could be directly scored by traditional algorithms, such as HITS [101] and PageRank [175]. In a survey with 40 subjects, each submitting one freely chosen query, EigenRumor outperformed a basic inlinks algorithm for 18 queries (45%), underperformed for 2 queries (5%), and tied for 19 queries (48%). For the remaining query, neither algorithm could improve upon a TF-IDF baseline [15]. Upon observation, they noted that EigenRumor and inlinks performed indistinctly for more generic queries (e.g., "baseball"), while EigenRumor was superior for more specific queries (e.g., "baseball ichiro").

Another approach to enrich the link structure of the blogosphere was proposed by Kritikopoulos et al. [108]. In their approach, a directed graph was initially created having blogs as nodes, and with a link between each pair b_i and b_j weighted by a linear combination of both explicit and implicit features $\vec{\phi}(b_i, b_j)$. As an explicit feature, they considered the total number of hyperlinks running between the posts of

the two blogs. As implicit features, they considered both the number of common tags and the number of common authors between blog posts. Their intuition was that blogs sharing many common tags (or topics) or many common authors would have similar interests, and hence should be connected. Based upon this enriched graph, they proposed a generalization of the PageRank algorithm [175], called BlogRank. This algorithm scored a blog b according to:

$$score_{BlogRank}(b) = (1 - \epsilon) + \epsilon \sum_{b_j \in \mathcal{I}_b} \vec{w} \cdot \vec{\phi}(b, b_j) \tag{4.30}$$

where $\mathcal{I}_b$ denotes the set of blogs linking to blog b, $\vec{w}$ and $\vec{\phi}(\bullet)$ are the weight and feature vectors for b and each of its linking blogs, respectively, and $\epsilon \in [0, 1]$ is a damping factor, set to 0.85. This algorithm differed from the original PageRank by weighting links non-uniformly, instead using the weighted score combining the explicit and implicit link features. To evaluate their proposed algorithm, they used a sample crawl of the blogosphere comprising 7.9 m posts from 1.5 m blogs. As a baseline, they used the standard PageRank. In a user study, they showed that BlogRank significantly outperformed PageRank, with an average gain of 250% in terms of a click-based success index [107].

A related problem to detecting influential bloggers was investigated by Kale et al. [94]. Specifically, they addressed the problem of detecting all blogs that a given blog would trust even if these blogs are not directly connected. To this end, their approach initially created *polar links* between directly connected blogs. A polar link was derived by weighting an existing hyperlink according to the sentiment polarity (positive, negative, or neutral) detected from the text surrounding the hyperlink (i.e., its anchor-text, expanded by a fixed length to the left and right directions). For polarity detection, they employed lexicons of positive ($\mathcal{L}_+$) and negative ($\mathcal{L}_-$) words, in a similar fashion to the lexicon-based opinion search approaches described in Section 3.2.1. In particular, the link between blogs b_i and b_j was weighted as:

$$w_{ij} = \frac{tf_{\mathcal{L}_+, v_i} - tf_{\mathcal{L}_-, v_i}}{tf_{\mathcal{L}_+, v_i} + tf_{\mathcal{L}_-, v_i}}, \tag{4.31}$$

where $tf_{\mathcal{L}_+, v_i}$ (resp. $tf_{\mathcal{L}_-, v_i}$) denoted the frequency of positive (resp. negative) lexicon terms in the context v_i surrounding the original hyperlink

from b_i to b_j. These initial polarity estimations formed a belief matrix, which was then used to spread the estimations to the entire blog graph through a series of iterative atomic propagations [75]. To test their approach, they used a sample of 300 political blogs labelled as either democratic or republican [1]. To detect the influence zone of bloggers with different political inclinations, they selected the top three most linked blogs from each of the democratic and republican communities as two different seed sets. The propagated influence from each seed set was then verified against the labelled inclinations of each influenced blog. Their results showed that accounting for the polarity of a link consistently improved the influence spreading mechanism compared to using the plain link structure.

Nallapati et al. [163] introduced a topic-sensitive approach to identify influential bloggers. Latent topic modelling had been previously used [43, 54] to model topic-sensitive word and link generation processes, using either Latent Dirichlet allocation (LDA [23]) or Probabilistic Latent Semantic Analysis (PLSA [85]). In their approach, Nallapati et al. combined LDA and PLSA to simultaneously model topic discovery and topic-sensitive influence. In particular, LDA was used to model the word and link generation processes of *linking* blog posts, while PLSA was employed to model the generation process for the words in the *linked* posts. As a result, the topic-sensitive influence of a blog b given a query q was estimated as:

$$\mathrm{P}(b|q) \propto \sum_{z \in \mathcal{Z}} \mathrm{P}(b|z)\,\mathrm{P}(q|z)\,\mathrm{P}(z), \tag{4.32}$$

where $z \in \mathcal{Z}$ is a latent topic, the probabilities $\mathrm{P}(b|z)$ and $\mathrm{P}(q|z)$ estimate the likelihood that the blog b and the query q are about the topic z, and $\mathrm{P}(z)$ denotes the prior probability of this topic in the underlying corpus. To evaluate their approach, they used a crawl of 8.3 m blog posts from July 2005, later pruned to leave out unlinked posts. The final subset of 2.2 k posts with at least two outgoing links and 1.8 k posts with at least two incoming links within the crawled corpus was modelled as a linking-linked bipartite graph. Through a cross-validation, they showed that their combined approach had a much higher predictive power in terms of log-likelihood compared to the approach using only LDA [54].

This observation held regardless of the target number of topics, attesting to the effectiveness of the proposed model for modelling topics and influence on the blogosphere.

4.5 Faceted Relevance

In its simplest form, blog search is concerned with the identification of blogs with a central and recurring interest on the topic of a query, as discussed in the previous sections. Nevertheless, the heterogeneous nature of the blogosphere reveals similarly relevant blogs with rather distinct characteristics, from different writing styles to different viewpoints on the same topic. One way of facilitating the access to the multiple facets of blog search results is through an exploratory search interface [83]. This interface would allow users to rapidly navigate to relevant results that fulfil a desired facet through a filtering mechanism [82]. In order to enable such a mechanism, a few approaches have been proposed to detect the extent to which a given blog satisfies a particular facet.

For instance, Liu et al. [126] proposed a classification approach to estimate the inclination of blog search results toward different facets. In their approach, each retrieved blog b was represented by two sets of blog post snippets: the set Υ_b of snippets published by the blog itself, and the set $\Upsilon_{\mathcal{I}_b}$ of snippets from blogs that linked to b. The former provided a sample of the blog's posted content, while the latter conveyed a sample of the blogosphere view on the blog. Each snippet $\upsilon \in \Upsilon_b \cup \Upsilon_{\mathcal{I}_b}$ consisted of the title and the first two or three sentences from a blog post. Using this representation, they conceived a two-stage classification approach. In the first stage, multiple SVM classifiers were used to predict the probability that a snippet belonged to different facets, with the 2,000 words with highest information gain [225] used as features. Formally, the probability $\mathrm{P}(c|\upsilon)$ that a snippet $\upsilon \in \Upsilon_b \cup \Upsilon_{\mathcal{I}_b}$ would match the facet c was estimated by fitting to a sigmoid [182] the output of the SVM_c classifier, trained for that particular facet:

$$\mathrm{P}(c|\upsilon) = \frac{1}{1 + \exp(\alpha\,\mathrm{SVM}_c(\phi(\upsilon)) + \beta)}, \tag{4.33}$$

where $\vec{\phi}(\upsilon)$ was the feature vector associated with the snippet υ, and the α and β parameters were obtained as maximum likelihood

estimates from the training data. In the second stage, blog features were generated by summarizing the probability that the snippets of a blog belonged to each facet. For this stage, different classifiers were used, including SVM and neural networks. For their experiments, a total of 4,4 k blogs listed for eight broad categories (art, business, education, health, law, politics, religion, and technology) were collected from public blog directories. To obtain posts published by and linking to each blog, they used the Google Blog Search API, resulting in a total of 86.5 k post snippets (an average of 19.6 snippets per blog). To evaluate the classification performance of the first stage, they propagated the known categories of blogs to their collected posts, so as to automatically obtain post-level ground-truth. While this procedure introduced some noise, they obtained a micro-level F1 of 53% over all eight categories. Despite the modest performance at the first stage, cross-validation results on the second stage yielded much better results, with up to 86% F1. Compared to a single-stage classification baseline that treated each blog as a "bag-of-snippets," they observed an improvement of 26%.

In contrast to the topically oriented facets considered in the previous approach, Gao et al. [60] proposed to identify latent attributes of a blog as facets. Following the experimentation paradigm provided by the faceted blog distillation task of the TREC 2009 Blog track [136], such attributes were defined as three groups of dual facet inclinations: opinionated versus factual, personal versus official, and in-depth versus shallow. The first group investigated the opinionated nature of the blogosphere as discussed in Section 3.2, however at the blog level. The second group sought to contrast personal blogs and those serving as the public outlet of an organization. Finally, the third group touched the language aspect of the blogosphere, by seeking to distinguish between bloggers with a thoughtful or a superficial writing style. To classify blogs into each of these facet inclinations, Gao et al. explored content features, such as average post or sentence length, or the occurrence of Internet slangs (e.g., "LOL") or emoticons (e.g., "=)"), and subjective lexicon features based upon SentiWordNet [55] and Wilson's compilation of subjective terms [222]. To weight the relative importance of each feature, they employed an SVM classifier for each facet inclination. In addition, they proposed to further improve the accuracy

of each classifier by expanding the initial set of features with feedback features, automatically extracted from blogs known to adhere to each facet inclination. Cross-validation results using the TREC 2009 Blog track queries showed that the basic model without feedback features performed comparably to the best performing systems at TREC 2009 [136] for some of the considered inclinations. Moreover, the feedback features resulted in significant improvements over the basic model for almost all inclinations.

Similarly, Gerani et al. [63] proposed a data-driven approach to identify personal blogs, in contrast to those written in an official role. They hypothesized that such blogs would more likely express opinions, and hence could be identified using opinion features. To obtain such features, they automatically extracted the highest ranked subjective words (according to a document-frequency-based mutual information metric) from all blog posts judged as opinionated in the TREC 2006, 2007, and 2008 Blog tracks [134, 171, 173]. To identify personal blogs, they first inferred the subjectiveness of each post p, by averaging the weights w_t of all subjective words $t \in p$:

$$score_{Personal}(p) = \frac{1}{l_p} \sum_{t \in p} w_t \, tf_{t,p}, \tag{4.34}$$

where $tf_{t,p}$ and l_p denote the frequency of the term t in the post p and the total length of p, respectively. Likewise, the personal inclination of each blog b was computed as the average subjectiveness of its posts:

$$score_{Personal}(b) = \frac{1}{n_b} \sum_{p \in b} score_{Personal}(p), \tag{4.35}$$

where n_b is the total number of posts in b. The final ranking of blogs was then produced by linearly combining the initial relevance ranking with the inferred personal ranking. In order to evaluate their proposed approach, they performed a cross-year validation using the query sets from the TREC 2009 and 2010 Blog track. On the TREC 2009 query set, their proposed approach significantly outperformed all three standard baselines provided by the TREC Blog track organizers [136]. However, on the TREC 2010 query set, the observed improvements were not significant, primarily due to the prevalence of harder queries, with much fewer relevant blogs.

4.6 Summary

This section has surveyed several approaches devoted to searching entire blogs (as opposed to blog posts) in response to a user's query. The described approaches covered a range of important aspects involved in tackling blog search, from the granularity issues involved when estimating the topical relevance of a blog or when performing relevance feedback, to other distinctive challenges posed by the structure of the blogosphere, such as its temporal and interconnected nature. Lastly, we addressed the task of identifying blogs fulfilling a desired facet, as a further step toward facilitating access to the information on the blogosphere. Together with the previous section, this section covered search tasks where the blogosphere was the target. In the next section, we survey cases where the blogosphere is instead used as a means to enable search tasks in other domains.

5

Blog-Aided Search

Search *within* the blogosphere, i.e., the search of blog posts or full blogs, are not the only use cases for the blogosphere in a search-oriented environment. In particular, the blogosphere is also used as a means to enable or enhance other search tasks, in contrast to being the target of the search. Indeed, applications include the identification of important news stories for display on news Web sites, the detection of topics and trends over time and the analysis of blog content for market analysis and brand monitoring. In this section, we examine non-blog search tasks that nonetheless can be enhanced through the use of blogs. Section 5.1 surveys the task of inferring the importance of news stories using the blogosphere. Section 5.2 details graph search engines and trend-detection approaches that leverage blogs to detect and present useful trends in blogging behaviour. Section 5.3 surveys the field of market analysis and brand monitoring using the blogosphere.

5.1 Inferring News Importance

The blogosphere is well-known as a medium for news reporting and discussion [140, 146, 206]. A poll by Technorati found that 30% of bloggers

considered that they were blogging about news-related topics [146]. Similarly, Mishne and de Rijke [156] showed a strong link between blog searches and recent news, as discussed in Section 2. Indeed, almost 20% of searches for blogs were news-related. As an illustration, Thelwall [206] explored how bloggers reacted to the London bombings, showing that bloggers respond quickly to news as it happens. Furthermore, both König et al. [106] and Sayyadi et al. [192] 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, newspaper editors perform the difficult task of deciding which stories are sufficiently newsworthy to place on the front or content pages of their (e-)newspaper. Indeed, it is of particular note that a newspaper's Internet portal is becoming increasingly important, as millions of users consult e-newspapers to find out the most interesting events and stories occurring worldwide [169]. Moreover, e-newspapers are editorially challenging, as they lack a fixed publication cycle and are instead updated and edited continuously with breaking news. Similarly, Web-based news aggregators (such as Google News[1] or Yahoo! News[2]) give users access to broad perspectives on the important news stories being reported, by grouping articles into coherent news events. Relatedly, very often in a given news article, some newspapers or news Web sites will provide links to related blog posts, often covering a diverse set of perspectives and opinions about the news story. These may be hand selected, or automatically identified.

However, the high volume and rate at which news content is currently created highlights the need for automatic tools to sort through the large volume of news in real-time, identifying the most currently newsworthy stories for display, i.e., as an electronic aid for editors. This editorial task can be seen as a ranking problem. For example, on the homepage of a news website, current news stories are ranked by their perceived newsworthiness at that time. Highly newsworthy stories receive a prominent placement on the page, while lesser stories are displayed less prominently or not at all. Figure 5.1 illustrates this for

[1] http://news.google.com/.
[2] http://news.yahoo.com/.

Fig. 5.1 An example of news story ranking on the BBC.co.uk Web site — 15/12/2011.

the homepage of the BBC.co.uk news Web site, where important stories are provided with a larger allocated story area and headline font.

It was proposed that the blogosphere, given its strong news focus and tendency to report and discuss recent events, would be a prime source of evidence from which to infer the importance of current news stories from a blogger's perspective [136]. The assumption is that the blogger population will be sufficiently similar to that of newswire's audience that the importance estimates obtained from the blogosphere will be a useful indicator for newspaper editors. To investigate whether this is the case, the *Top Stories Identification task* was devised within the Text REtrieval Conference (TREC), with the purpose of exploring how accurately top news could be identified using the blogosphere [174].

The identification of top news stories from one or more news providers can be considered as answering the question "*what are the most important news stories today.*" This can be seen as a ranking task, where a stream of current news stories — either pre-provided by major news providers or incrementally crawled online — are ranked by their newsworthiness for placement on the homepage or category pages of a news website using evidence from the blogosphere. In particular, for a given day d and a set of news stories $\mathcal{C}$, represented by news articles, rank all stories in $s \in \mathcal{C}$ by their importance for day d using only evidence from blog posts, i.e., rank by $score(d, s)$.

Intuitively, the importance of a news story can be inferred by measuring the volume and intensity of discussion about that story. The more a story is discussed, the more important it is considered to be. In the case of the blogosphere, this can be estimated using the number of blogs or individual blog posts that discuss a news story. Indeed, based upon this initial idea, strategies of varying complexity and effectiveness have been proposed in the literature.

The simplest means to estimate a news story's importance is to count the number of blog posts that directly cite the story in question. Mejova et al. [147] considered hyperlinks pointing to a news article discussing the story to be a citation and counted the number of blog posts containing such citations to form a score for each story. Formally, the importance of a news story s in this approach can be described as:

$$score(d, s) = \sum_{p \in \mathcal{D}^{(d)}} \mathbf{1}_{\mathcal{I}_s}(p), \tag{5.1}$$

where s is a news story, d is the day of interest, $\mathcal{D}^{(d)}$ is the set of documents (blog posts) published on day d, p is a post in $\mathcal{D}^{(d)}$ and $\mathbf{1}_{\mathcal{I}_s}(p)$ is the indicator function, which determines whether the post p is among the posts $\mathcal{I}_s$ linking to s. This is a high-precision approach, which uses only the strong relationship that a direct hyperlink provides to relate blog posts and news stories. In practice, when tested during TREC 2009, this strategy provided limited effectiveness due to the sparsity of links to each news story within the blogosphere.

More effective strategies make use of the textual similarity between a news story and blog posts to estimate the number of blog posts

discussing the story. For instance, Xu et al. [223] estimated the current newsworthiness of a story by summing the BM25 scores for each blog post that was published in the prior 24 hours to that story, thereby using a traditional information retrieval model to estimate relatedness between a story and recent blog posts:

$$score(d, s) = \sum_{p \in \mathcal{D}_s^{(d)}} score_{BM25}(s, p), \tag{5.2}$$

where $\mathcal{D}_s^{(d)}$ is the set of blog posts matched to the story s from the set of all blog posts published on day d. Indeed, this approach achieved the third best performance during TREC 2010 [174].

Similarly, Lin et al. [121] textually compared news stories and blog posts to estimate the relevance of posts relating to a news story. In particular, they used one vector-space representation to describe each news story. Each blog post published on day d was also represented in vector space. These vectors were defined by the term frequency inverse document frequency (TF-IDF) scores of each term. The importance of a news story on day d was estimated as the summation of the similarity scores between that story's vector and each blog post vector. Similarity was defined using the cosine measure:

$$score(d, s) = \sum_{p \in \mathcal{D}^{(d)}} sim(s, p), \tag{5.3}$$

where $\mathcal{D}^{(d)}$ is the set of all blog posts published on day d, $sim(s, p)$ returns the cosine similarity between the story s and the post p, where both s and p are represented as TF-IDF vectors. This was the second most effective approach at TREC 2010 [174].

Another approach proposed by McCreadie et al. [143] exploited the textual similarity between a news story and blog posts in a different manner. In particular, they proposed to model a story's newsworthiness as a voting process [137]. Under this strategy, blog posts were ranked for each story using a traditional information retrieval weighting model. Based on the number and relatedness of retrieved posts to each story, the importance of each headline on that day was inferred. Under this approach, the importance of a single news story is estimated as:

$$score(d, s) = \sum_{p \in (\mathcal{D}^{(d)} \cap \mathcal{D}_s)} score(s, p), \tag{5.4}$$

where $\mathcal{D}^{(d)} \cap \mathcal{D}_s$ is the set of only those blog posts that were both retrieved for the story s from the set of all posts (regardless of date) $\mathcal{D}_s$ and are from the set of blog posts published on day d, $\mathcal{D}^{(d)}$. Critically, this approach differs from that of Xu et al. [223] (see Equation (5.2)) in two ways. Firstly, the number of retrieved blog posts in $\mathcal{D}_s$ was fixed to 1000, hence only highly scored posts are considered. Secondly, by counting only those posts in $\mathcal{D}^{(d)} \cap \mathcal{D}_s$, McCreadie et al.'s approach implicitly contrasts the volume of posts returned on day d to those returned on days other than d. The intuition was that a story should be important on day d if the majority of the 1000 top retrieved posts were published on day d. To retrieve the 1000 blog posts in $\mathcal{D}_s$, McCreadie et al. used the parameter-free document weighting model DPH from the Divergence from Randomness framework [7]. This approach was the best performing run submitted to TREC 2009 [136] and was later shown to be effective on the TREC 2010 dataset [142].

Additionally, by examining the historical importance of a headline over time, McCreadie et al. later found that marked improvements in effectiveness could be achieved on the TREC 2009 dataset [141]. In particular, they weighted the score for each blog post in $\mathcal{D}_s$ by the time elapsed between the publication of those posts and the query day d. For each day worth of elapsed time since the query day d, a different weight was calculated using a Gaussian curve distribution as follows:

$$score(d,s) = \sum_{p \in \mathcal{D}_s} Gauss(\Delta p)\, score(s,p), \tag{5.5}$$

where Δp is the time elapsed since the publication of p and the query day d (in days). $Gauss(\Delta p)$ calculates a weight for p based upon Δp as follows:

$$Gauss(\Delta p) = \frac{1}{\omega\sqrt{2\pi}} \exp \frac{-(\Delta p)^2}{(2\omega)^2}, \tag{5.6}$$

where the parameter ω defines the width of the Gaussian curve. By increasing the value of ω, posts with a larger Δp are considered. McCreadie et al. found that a ω value of approximately 1 — that considers only posts made within a time window of at most three days of d [141] — was the most effective.

News story ranking has also been approached via probabilistic language modelling. In this case, the likelihood of each story generating

an aggregate language model of recent blog posts indicates the story's newsworthiness. One such approach was proposed by Lee et al. [114], whereby clustering was used to create multiple topic models. These topic models were then compared to a story model generated from the top retrieved blog posts for that story. In its simplest form, a score for each news story was then calculated as:

$$score(d, s) = \mathrm{P}(\mathcal{D}^{(d)}|s)\,\mathrm{P}(s), \tag{5.7}$$

where $\mathrm{P}(\mathcal{D}^{(d)}|s)$ is the probability of a set of blog posts $\mathcal{D}^{(d)}$ published on day d given the news story s, and $\mathrm{P}(s)$ is a prior probability for the story s. In practice, $\mathrm{P}(\mathcal{D}^{(d)}|s)$ was estimated as the combination of a query language model and news story model. This defined how related the story was to the current day. The stronger the relation, the more important the news story was considered to be. $\mathrm{P}(s)$ incorporated the prior informativeness of the query terms and also the temporal profile of the query [114]. This was the best performing run submitted to TREC 2010.

Later, McCreadie et al. [142] proposed an aggregate learning to rank strategy [128] to more accurately estimate the importance of a news story. In particular, they proposed a framework whereby multiple prior approaches were leveraged to provide story importance estimates given one or more textual representations of a story and some temporal constraints. Under this approach, different importance estimates $score(d, s)$ were used as features and combined using learning to rank. The framework combined three components, namely: a story ranking model, a day for which to rank, and a representation of the story s. By combining one of each of these three, a story ranking feature was generated. McCreadie et al. used multiple prior story ranking approaches [141, 223], generated different query representations using collection enrichment [130], and estimated importance for multiple dates around d, to create a variety of features for story ranking. When tested upon both the TREC 2009 and 2010 datasets, this approach was shown to markedly outperform the individual approaches upon which it was built, which included top systems from TREC 2009 and 2010 described earlier.

In summary, the estimation of news story importance using the blogosphere has become an active research area, tackled by groups

worldwide. This task has raised interesting issues regarding how to identify the most related blog posts to a story, how to aggregate them for news story scoring, and how to best model the timeliness aspects of blog reporting/discussion within a story ranking system. Indeed, each of these are still open research directions. As we will discuss in Section 7, a similar task can be envisaged within a microblogging environment, using tweets to detect emerging news stories in real-time (e.g., Petrović et al. [180] have tackled event detection in tweet streams).

5.2 Trend Detection

An alternative popular application of the blogosphere is trend detection. A trend can be considered to be a pattern of behavior over time. In a blogging context, there are a variety of trends that one might be interested in. For example, a user might be interested in the most frequently cited Web pages or news articles by bloggers [68], while marketing groups might be interested in recent trends in brand and product discussions in the blogosphere [22, 86]. In general, trend detection is particularly useful for enabling temporally oriented search tasks, such as real-time search [49], and is used by graph search engines, such as IceRocket,[3] to display trends in blogging behavior.

A variety of academic graph search engines have been developed to capture and visualise trends within the blogosphere. One of the earliest is BlogPulse,[4] developed by Glance et al. [68]. BlogPulse tracks n-gram frequencies in blogs over time to identify key phrases that describe topics that are currently of interest. In particular, it identifies key n-grams in the blog post corpus by their popularity and then filters them using a set of heuristics to find only the most informative ones. These initial n-grams are subsequently clustered into separate topics. For each resulting topic, the component n-grams are used to search the blogosphere for a single paragraph for display to the user. BlogPulse also supports named entity tracking, e.g., people or products, using similar n-gram frequency-based methods.

[3] http://trend.icerocket.com/.
[4] http://www.blogpulse.com/.

Later, Chi et al. [39] noted that the output from purely n-gram frequency based methods could be noisy, as relatively few unknown bloggers could generate a false trend. To counteract this, they proposed two alternative approaches based on singular value decomposition (SVD), Scalar Eigen-Trend and Structural Eigen-Trend. These approaches combine both frequency-based trendiness with a measure of how authoritative a blogger is. In particular, Scalar Eigen-Trend records the contribution of bloggers over time to different trends: the more trends they contribute to, the more authoritative they are considered to be. Instead, Structural Eigen-Trend uses a variant of the HITS algorithm [100] to measure the blogger authority in terms of hubs and authorities. On a blog dataset, Chi et al. show that using these SVD-based approaches, interesting trends that are not apparent using pure n-gram frequency-based approaches could be detected.

Recently, Schirru et al. [193] leveraged trend detection at a blog post level rather than n-gram level in order to create an application for the semi-automatic analysis of topics and trends for expert users. In particular, they built a unigram-based system that tracks trends over time. Latent semantic analysis (LSA) was used to reduce noise in the set of unigrams. The degree to which a term is trending was estimated using the term frequency inverse document frequency (TF-IDF) term weighting scheme. Only blog posts from within a 4 day time-window were used to detect trending terms. Blog posts that contain trending terms were then clustered into topics. The composition of a topic was then be exposed based on the number of posts that contain each trending term, in addition to the distribution of related blog posts over time.

Overall, trend detection approaches that leverage the blogosphere have proved to be popular, with graph search engines like BlogPulse receiving over 60,000 requests each day.[5] Current approaches, such as those by Chi et al. [39] and [193], monitor term n-gram frequencies in blog posts to identify and track topics and brands over time. In the next section, we examine a group of trend detection techniques that track brands and products.

[5] http://www.freewebsitereport.org/www.blogpulse.com.

5.3 Market Analysis

Market analysis [22] is a major application for a specific group of trend detection techniques. These techniques focus on tracking specific brands or products, normally on behalf of large companies, so that they can monitor product popularity, gain implicit product feedback or identify groups for targeted advertising. In contrast to generic trend analysis techniques, market analysis normally starts with a user query, relating to a specific topic, product or brand that the searcher wants to examine. Companies are well-known to use the blogosphere as a tool to monitor their brands and products. Indeed, there are many companies which provide such services, e.g., Nielsen.[6] Moreover, as companies spend more of their marketing budget on social media campaigns,[7] effective market analysis and trend monitoring tools become critical to determine the value added by such campaigns. For example, market analysis tools were used to analyse the effect of the Skittles social media brand awareness campaign in 2011.[8]

One of the most prominent and general market analysis tools is the WebFountain analytics engine implemented by IBM [73]. WebFountain is a general analytics engine that provides large-scale parallel access to continually updating unstructured data, like blogs. It can be used as a market analysis tool through the use of specialist data miners, which perform tasks such as statistics aggregation, trend tracking, relationship extraction, and clustering.

Various approaches for the exploration of topics or brands have been proposed. For instance, Ziegler et al. [241] developed tools specifically for market analysis using a variety of data sources, including blogs. In particular, they focused on relation discovery between entities, e.g., people and brands, based on frequently co-occurring n-grams. To this end, they built an co-occurrence graph describing the relations between n-grams. They also proposed an approach to compare different brands based upon their semantic profile. In particular, a profile for an entity was built by retrieving highly relevant documents to that entity and

[6] http://www.nielsen-online.com/products_buzz.jsp?section=pro_buzz.
[7] http://www.time.com/time/business/article/0{,}8599{,}1958400{,}00.html.
[8] http://socialwoot.com/brands/skittles-social-media-campaign.html.

categorizing those documents to form a weighted category graph using the DMOZ Open Directory Project.[9] The authors transformed such graphs into a vector-space representation such that different entities could be compared.

There has also been prior work examining large-scale sentiment analysis in blogs [178], i.e., the classification of discussions relating to brands or products as positive or negative. Such approaches can be used to gauge the public reaction to new products as they are released. For example, Pimenta et al. [181] investigated two blog post sentiment analysis approaches. Firstly, they built a sentiment classifier using the SentiWordNet [55] opinion mining toolkit. Secondly, they identified reoccurring lexical patterns from the blog posts that mark the expression of an opinion. These approaches leveraged part-of-speech tagging of the corpus to identify adjectives and adverbs in documents that were subsequently classified as containing positive or negative sentiment indicators [211]. Indeed, such opinion mining approaches hold strong similarities to opinion search in blog posts, as described previously in Section 3.2. The difference in this case, is that the output of the system is not a list of blog posts with a particular sentiment, but rather an aggregation of those sentiments over time into a global trend.

In summary, market analysis and trend monitoring are vital tools for companies as they spend ever increasing proportions of their marketing budgets on social media campaigns. Scalable tools like IBM's WebFountain [73] have been developed to analyze large datasets, while current research is focused on accurately detecting sentiments within blogs at scale [69]. Market analysis will likely remain an active research area for the foreseeable future, with a strong focus on tracking brand awareness and sentiment over multiple streams of social media data [241]. Moreover, emerging data sources like Twitter introduce new and interesting problems for sentiment detection in short texts [176].

5.4 Summary

This section has surveyed tasks and approaches where the blogosphere is used to aid or enhance search tasks other than pure blog search.

[9] http://www.dmoz.org.

In Section 5.1, we examined the task of inferring the importance of news stories using the blogosphere. We showed how the TREC Top Stories Identification task focused research into news within blogs and highlighted effective approaches. Meanwhile, in Section 5.2, we surveyed the field of trend detection approaches that leverage blogs to detect and present useful trends in blogging behavior. Section 5.3 examined approaches and tools for the specific task of market analysis and brand monitoring using the blogosphere. Overall, accurate and timely story ranking and trend-detection techniques using the blogosphere continue to constitute promising research areas with industry interest and impact for prominent companies, e.g., Thomson Reuters and Nielsen Media Research. Such techniques will likely continue to be of interest as the number, size and reach of social media sources, like Twitter, increases.

6

Publicly Available Resources

Thus far, we have described a wide range of search tasks and techniques for the blogosphere. This section surveys the resources that were developed over the years to facilitate experimental research within a blogosphere context. Such resources encapsulate shared data, corpora, and test collections for evaluation. In particular, shared test collections are paramount for evaluating progress with respect to the state-of-the-art, and focusing researchers' interest on specific common problems. The section ends with a discussion of the major forums where researchers can disseminate their findings on searching the blogosphere.

In general, different information needs usually demand different information retrieval techniques which, in turn, would benefit from being evaluated on a suitable standard benchmark. This is the overall idea of the Text REtrieval Conference (TREC), one of the major forums for research in information retrieval [216]. TREC was introduced in 1992 in a co-sponsorship between two U.S. government agencies: the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA).

TREC can be seen as a modern instantiation of the so-called Cranfield experimentation paradigm [42]. Its overall goal is to support information retrieval research by providing the necessary infrastructure for the evaluation of retrieval techniques on a common benchmark, called a *test collection*. A test collection comprises three components: a set of documents, a set of stated queries representing specific information needs (called *topics*), and a set of relevance assessments, which function as a mapping between each topic and the documents deemed as relevant for this topic.

Since its inception in 1992, TREC has witnessed a remarkable increase in the number of participating groups working on various different retrieval scenarios (known as *tracks* in the TREC jargon). Tracks have been created for different types of document corpora (e.g., Web pages, enterprise documents, or chemical patents), or different retrieval tasks (e.g., relevance feedback, filtering). We refer the reader to [217] for an overview of the first 15 years of TREC.

The Blog track which ran as part of TREC 2006–2010, represents one of three forays TREC has made into the searching of "personal" documents. In particular, a blog post often represents a viewpoint of an individual, and in this way differs from the other document corpora used at TREC, such as general Web crawls, newswire, and patents. Apart from the Blog track, the Enterprise track, through its e-mail known item (2005) and discussion (2006) search tasks [44], and more recently the Microblog track (2011) [172] have dealt with search engine retrieval of personal documents.

Of course, the test collections generated as part of the TREC Blog track do not represent the only datasets available for research on the blogosphere. For instance, BlogPulse and later Spinn3r created samples of the blogosphere for the purposes of the Workshop on the Weblogging Ecosystem (WWE) series, and its successor, the International Conference on Weblogging and Social Media (ICWSM), through their data challenge. In the remainder of this section, we discuss the blogosphere samples and corresponding evaluation tasks that were formed from the TREC Blog track (Section 6.1), as well as the ICWSM datasets (Section 6.2). Lastly, we highlight other salient resources that are worthy of investigation for research on the blogosphere (Section 6.3).

6.1 TREC Blog Collections

A corpus is a fundamental component of a test collection for information retrieval research. For the Blog track, a large common sample of the blogosphere was required, which could be used by all users of the test collections. This required a large sample of blog posts collected over a substantial time period, representing many different blogs. For the TREC 2006–2008 campaigns, the organizers of the track created the Blogs06 corpus [131]. This corpus is described in further detail in Section 6.1.1. The Blogs08 corpus, which was both larger in volume and timespan, was introduced for TREC 2009 [136], and is detailed in Section 6.1.2.

6.1.1 The TREC 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 148 GB in size, with three main components consisting of 38.6 GB of XML feeds (i.e., the blog), 88.8 GB of permalink documents (i.e., a single blog post and all its associated comments) and 28.8 GB of HTML homepages (i.e., the main entry to the blog). In order to ensure that the Blog track experiments can be 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. The organizers also endeavored 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 large blog hosting providers (e.g., Blogspot.com, Wordpress) are not accessed 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 might 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 148 GB. The number of permalink documents in the collection amounts to over 3.2 million, while the number of feeds is over 100,000 blogs. Macdonald and Ounis [131] documented the development methodology and the statistics of the collection. Some salient statistics of the Blogs06 corpus are listed in Table 6.1. The corpus was used for the Blog track for three years (2006–2008).

6.1.2 The TREC Blogs08 Corpus

To facilitate research into the blog search task, requiring views of the evolving nature of blogs across a large timespan, the need for a new and larger blog corpus was identified during the TREC 2007 campaign. To this end, the organizers started the creation of the Blogs08

Table 6.1. Statistics of the Blogs06 and Blogs08 test collections.

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	25 GB	453 GB
Total uncompressed size	148 GB	2,309 GB
Feeds (uncompressed)	38.6 GB	808 GB
Permalinks (uncompressed)	88.8 GB	1,445 GB
Homepages (uncompressed)	20.8 GB	56 GB

corpus in late 2007. In particular, the desired properties were more feeds (blogs), more blog posts, collected over a significantly longer time period than the 11 weeks of Blogs06. Macdonald et al. [136] documented the methodology for creating the corpus and its statistics. Firstly, in addition to the feeds in Blogs06, they collected more feeds by sampling from online blog directories and from the recent updates list of major blog hosting providers (e.g., Blogspot and Wordpress). Similarly to Blogs06, they also used blog search engines to search for blogs of interest to a general audience. Finally, they used outgoing links from blogs in Blogs06 to identify further blogs. Over one million blogs were identified using these processes. Note that, in contrast to Blogs06, spam blog feeds were not targeted to be added into Blogs08 — however, it is highly likely that the collection does contain some. Indeed, the recent updates list obtained from major blog hosting providers was found to contain much spam, similar to the splog presence in the ping servers identified by Kolari et al. [104].

According to Macdonald et al. [136], the crawling strategy for Blogs08 was very similar to that used for Blogs06. One small difference was that blog homepages were only collected once, rather than each week. They monitored the one million blogs on a weekly basis from 14th January, 2008 to 10th February, 2009. This timespan of over one year allowed a substantial sample of the blogosphere to be obtained and facilitated 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, the TREC 2008 Blog track continued to use Blogs06, 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 6.1. Both Blogs06 and Blogs08 continue to be distributed by the University of Glasgow.[1]

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

6.1.3 What can be Evaluated Through TREC?

During the first three years of the Blog track [171, 134, 173], four different search tasks were investigated. These include three post retrieval tasks, in which the retrieval unit was a blog post, and one blog search task, in which entire blogs were the retrieval units. In 2009 and 2010, the track featured a different post retrieval task, in addition to a refined version of the blog search task investigated until then. Table 6.2 displays all these tasks along the timeline of the TREC Blog track. A description of each of them is given in the remainder of this section. Guidelines from the various TREC Blog track campaigns (2006–2010) are archived on the Blog track wiki.[2]

Opinion Finding Task This task was the first and longest-lived among all tasks investigated in the first three editions of the TREC Blog track. Placed in the broader area of sentiment analysis [178], it is motivated by the knowledge that can be extracted when bloggers are considered collectively. In other words, it aims to find out what people think about a particular topic. Following the taxonomy described in Section 2.3, it impersonates the information needs expressed mostly as

Table 6.2. Timeline representation of the retrieval tasks addressed in the Blog track from 2006 to 2010. Opinion finding (OF), polarity (PL), baseline *ad hoc* (BL), and top stories identification (TS) are post retrieval tasks, whereas blog distillation (BD) is a blog retrieval task. We also link to the relevant section of this survey where approaches for each task are described, as well as the corpus used and the appropriate overview paper for that year.

task	Section	2006	2007	2008	2009	2010
OF	Section 3	✓	✓	✓		
PL	Section 3		✓	✓		
BL	Section 3			✓		
TS	Section 5				✓	✓
BD	Section 4		✓	✓	✓	✓
Corpus	Section 6	Blogs06	Blogs06	Blogs06	Blogs08	Blogs08
Overview paper	—	[171]	[134]	[173]	[136]	[174]

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

context queries, and can be stated as

> *find me blog posts with an expressed opinion about entity x*

where x represents a named entity, e.g., a person, an organization, an event, a location, etc.

Each participating system in this task was evaluated using a set of topics and their associated relevance assessments. Along the three years in which this task ran, three topic sets were created, with a total of 150 topics with corresponding relevance assessments. An example topic is shown in Figure 6.1.

Besides a unique identifier (a topic *number*), each topic represents a specific information need using three fields: *title*, a short representation of the information need, typical of queries submitted to Web search engines; *description*, a more verbose specification of the query intent; and *narrative*, a complementary specification of the underlying information need, usually including some examples of what should be considered relevant or irrelevant to this information need.

As the aim of this task is to find not only relevant content for a given topic representing a named entity, but also an expressed opinion toward this entity, the relevance assessment procedure for the posts retrieved for each topic has two levels [171]. The first level assesses whether a given blog post contains information about the target and is therefore relevant. The second level assesses the opinionated nature of the blog

```
<top>
  <num> Number: 1030 </num>
  <title> System of a Down </title>
  <desc> Description:
    What do people think about the metal band System of a Down and
    their music?
  </desc>
  <narr> Narrative:
    Any positive or negative comment about System of a Down,
    their music, albums, or songs is relevant. Opinions concerning
    the performers' personal lives or endorsements are not relevant.
  </narr>
</top>
```

Fig. 6.1 TREC 2008 Blog track opinion-finding task, topic 1030.

post,[3] if it was deemed relevant in the first assessment level. The inclination of the manifested opinions — as to whether they represent a negative or positive attitude toward the entity in question — was not effectively investigated until the second edition of the Blog track, as described next. This task has generated much interest in the research community, as exemplified by the large volume of research surveyed in Section 3.2. In TREC 2008, as will be described below, the task required the participating groups to deploy their opinion finding approaches on top of common standard topic relevance baselines, so as to assess the performance of the opinion finding techniques in multiple settings.

Polarity Task The polarity task was first introduced in 2007 as a natural addition to the opinion finding task [134], again in relation to context queries. Initially, this task was seen as a classification task, in which the inclination of a given manifested opinion should be determined. This perspective was changed in its subsequent edition — so as to simplify the evaluation procedure [173] — and the task began to be seen as a combined retrieval task, in which the participating systems should rank posts of different inclinations separately (i.e., to produce a "negative" and a "positive" ranking of posts). The topics for this task were the same as those used for the opinion finding task. Similarly to that task, the polarity task can be stated as

> *find me negatively (or positively) opinionated blog posts about entity x*

Based on this task definition, and the developed test collection, a number of approaches have recently appeared in the literature to effectively address this task, which have been summarized in Section 3.2.3.

Baseline Adhoc Retrieval Task After the first edition of the opinion finding task within the TREC 2006 Blog track, it was observed that most participants approached opinion finding as a two-stage process [171]. In the first stage — often referred to as the *topic relevance*

[3] According to the Blog track guidelines, the relevant or opinionated portion of a given post can appear either in the post itself or in one of the comments to the post.

baseline — their system applied traditional information retrieval techniques to retrieve posts likely to be relevant to a given topic, regardless of whether they expressed an opinion about this topic. Specific opinion finding techniques were then applied in a second stage, in order to rerank the posts retrieved in the first stage and place opinionated posts ahead of purely factual ones. The effectiveness of applying a particular opinion finding technique on top of a topic relevance baseline, however, could not be assessed directly, as the retrieval performance of the latter was not reported. Additionally, the comparison of the effectiveness of the opinion finding techniques deployed by different participants was not possible, as they were not applied to a common baseline.

In order to address the first issue, the TREC 2007 Blog track required the participants to also submit a run with the ranking produced by their topic relevance baseline — a ranking produced without using any specific opinion finding technique [134]. For also solving the second issue, in 2008, the Blog track introduced a separate baseline *ad hoc* retrieval task, using the same topics as the opinion finding task, however aimed at assessing the effectiveness of the participants' approaches at retrieving relevant posts regardless of their opinionatedness [173]. It can be stated as

> *find me blog posts about entity x*

Out of all runs submitted to this task, five strongly performing ones were chosen by the TREC organizers to form a common ground for the evaluation and direct comparison of the opinion finding component of the participants' approaches. Additionally, as the Blog track experience has shown that it is particularly difficult to outperform a strong topic-relevance baseline — in some cases, not even an ideal opinion finding technique would be able to do so [129] — this common ground also enabled the evaluation of the robustness of individual opinion-finding approaches across a range of statistically different baselines. The five standard baselines have been used widely in the opinion finding literature and in many of the approaches described in Section 3.2.

Top Stories Identification Task The top stories identification task was first introduced in the TREC 2009 Blog track, in order to

address another unique aspect of the blogosphere, namely, its reactive behavior to real-world events. Top stories identification is a form of blog-aided search, where the blogosphere is used to detect newsworthy events (see Section 5.1). Indeed, the blogosphere responds to events as they happen, by generating discussions and opinions [206], as well as news-related (context) queries to blog search engines [156] (see Section 2.3).

The general idea of the task as instantiated within the TREC Blog track is as follows: given a particular date, the participating groups should identify the top stories (i.e., news headlines) that emerged on that date [136]. It can be summarised as

find me the top news stories that emerged on date x

Participants were provided with a corpus of news headlines. These headlines should be ranked with respect to the date topic based on supporting evidence identified from the blog corpus. Each ranked news headline will be assessed based on how important it is within that day's news. In particular, two news headline corpora were used: articles kindly donated by the New York Times (TREC 2009); and the TRC2 corpus of newswire, kindly donated by Thomson Reuters (TREC 2010). The task was refined for TREC 2010, by asking for the top news in different news categories: world, United States, sport, science/technology and business [174]. Additionally, in a second stage, systems were also assessed at retrieving a diverse set of blog posts related to the news headline. A number of representative approaches have been described in Section 5.1.

Blog Distillation Task Differently from all the other tasks investigated in the TREC Blog track, the blog distillation task [134] is the only one dedicated to blog retrieval — i.e., instead of blog posts, the retrieval units in this task are entire blogs. A classical example for this task is a user looking for blogs to add to his or her feed aggregator so as to follow them on a regular basis. It can be stated as

find me blogs principally and recurrently devoted to concept x

```
<top>
  <num> Number: 1069 </num>
  <title> whisky scotch </title>
  <desc> Description:
    Find blogs that discuss Scotch whisky.
  </desc>
  <narr> Narrative:
    Scottish whisky (scotch) is an alcoholic spirit, made in Scotland and
    exported all over the world. Related issues include new malts, blends,
    distilleries and master blenders.
    Relevant blogs will regularly discuss issues relating to Scottish whisky
    (scotch) regularly.
  </narr>
</top>
```

Fig. 6.2 TREC 2008 Blog track blog distillation task, topic 1069.

where x represents a general concept, making this task a direct instantiation of the information needs expressed as *concept queries*, as defined in Section 2.3. An example topic is shown in Figure 6.2.

Strictly speaking, blog distillation queries differ from concept queries in that x is not required to be a high level concept, such as "politics" or "sports." In fact, it could even represent a named entity, as targeted by context queries. A major difference from context queries, however, is that the blog distillation task is primarily interested in informative content about x, regardless of whether this content contains an expressed opinion toward x. Indeed, in its first two editions, the task focused solely on the topical relevance aspect of blogs, aiming to retrieve blogs that publish mostly about a given topic, and that cover the topic across most of their published posts. In 2009, however, inspired by a position paper by [82], the blog distillation task was refined in order to consider the quality aspect of blogs [136]. In this new formulation, called *faceted blog distillation*, the task can be stated as

> *find me* ***quality*** *blogs principally and recurrently devoted to concept* x

where "quality" can be defined differently based on the specific "facet" of interest. For instance, one might be interested in blogs written by females, or in those written by experts. The facet inclinations used for the instantiation of this task in TREC 2009 and 2010 were opinionated

versus factual, in-depth versus shallow; and personal versus official [136, 174]. Approaches for blog search, including faceted blog distillation are described in Section 4.

6.2 ICWSM Data Challenge Corpora

The AAAI International Conferences on Weblogs and Social Media (ICWSM) is an interdisciplinary conference bringing together computer and social scientists on the subject of social media. ICWSM grew out of two events: an annual series of Workshops on the Weblogging Ecosystem (WWE) held in 2004–2006 in conjunction with the International World Wide Web Conference and the Spring Symposium organised by the American Association for Artificial Intelligence (AAAI) on Computational Approaches to Analyzing Weblogs (CAAW 2006).

The WWE 2006 workshop featured a data challenge "to encourage the use of this data to focus the various views and analyses of the blogosphere over a common space."[4] The dataset, provided by BlogPulse (then part of Intelliseek), comprised 10 million blog posts from one million blogs over three weeks in July 2005. This collection was restricted to only be used up to May 2006 and is thus no longer available. Three papers out of the twelve presented at the workshop [122, 170, 206] used the dataset.

ICWSM 2007 also featured a data challenge with a dataset with a limited lifetime. The dataset, provided again by BlogPulse (later part of Nielsen-Buzzmetrics, now NM Incite's My BuzzMetrics; hence this dataset is sometimes referred to as the Buzzmetrics dataset), contained about 14 million posts from 3 million blogs in May 2006. Shi et al. [196] compare this dataset to the TREC Blogs06 dataset in terms of topology and overlap. Four other papers [12, 72, 94, 167] used the dataset; an equal number of ICWSM 2007 papers made use of the TREC Blogs06 collection, and still other papers collected their own datasets which were not released.

Starting after 2008, ICWSM decided to pursue a broader data agenda, namely to assemble datasets for the social media research

[4] See the WWE 2006 Call for Papers at http://www.blogpulse.com/www2006-workshop/cfp.html.

community, to make them available under relatively few usage restrictions, and to be a central gathering place for data. For ICWSM 2009, a larger collection of 44 million blog posts from between August 1st and October 1st, 2008 was obtained from Spinn3r [29]. This collection was promoted through a workshop at the conference specifically for work using the 2009 dataset. The five papers presented [26, 35, 70, 200, 209] were largely descriptive of the dataset, but served as a longitudinal data point for authors to verify or contrast results from previous datasets. The ICWSM 2010 data challenge workshop reused the 2009 dataset, but additionally invited work on other publicly-accessible datasets and encouraged the sharing of data derived from the 2009 dataset. Two papers in that workshop [57, 148] used the ICWSM 2009 data.

For the 2011 conference, ICWSM collected a new, still larger dataset from Spinn3r [30]. This dataset contains 286 million items, including 133 million blog posts and 231 million social media publications (mainstream news, forum, and review items, etc) from the period of January 13th to February 14th, 2011. The sheer size of this collection (2.1 TB compressed) has created many challenges for the community, and not much work has yet been done with it as a result.

6.3 Other Resources

Not only test collections are useful for developing search engines for the blogosphere. Instead, resources in the form of data can be used to deploy or refine blog search techniques. For example, some opinion finding techniques rely on the use of dedicated dictionaries that contain terms indicative of opinionated content. In the following, we describe several resources that have been used to conduct research on blog search.

6.3.1 SentiWordNet

While some of the approaches developed for opinion finding on the blogosphere (e.g., [161], described in Section 3.2.1.1) have mined their own lexicons of opinionated words, the SentiWordNet lexicon [14, 55] has seen widespread use in the literature.

In particular, SentiWordNet defines two aspects of a word: its subjectivity (how likely it is to represent an opinionated versus objective statement), and its polarity (whether it has positive or negative connotations). In particular, each synonym set (synset) from WordNet [153] are scored by a committee of classifiers in a semi-supervised approach [55]. The later 3.0 version of SentiWordNet builds on this by using an iterative random walk process to improve accuracy. Using these approaches, word such as "good," "goodness" and "inspired" are found to be the most positive, while "abject," "deplorable" and "lamentable" are among the most negative.

6.3.2 Amazon Review Data

Many approaches for identifying sentiment or opinionated material require training data with which to learn. While objective text is relatively easy to find (e.g., newswire, Wikipedia), access to voluminous sources of opinionated text is less easy. To this end, Jindal et al. [93] identified the presence of a large amount of opinionated content within product reviews (e.g., films, products, hotels). They created a dataset of opinionated text by crawling product reviews from the Amazon e-retailer. In particular, the Amazon Product Review Data[5] contains 5.8 million reviews of books, music, DVDs, and other categories, from 2.14 million reviewers across 6.7 million products.

Since its release, many opinion identification and sentiment analysis techniques have used this dataset, which has shown good transferability to other domains, such as opinion identification in blogs [194] (see Section 3.2.1.2) or mining hotel reviews [177].

6.3.3 Newswire Corpora

The TREC 2009 and 2010 Blog track Top Stories Identification task [174] aimed to investigate the ranking of news stories by their importance at specific points in time using the Blogosphere. Importantly, news stories were kindly provided by two different news providers, namely: the New York Times and Thomson Reuters, from

[5] Available from http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.

Table 6.3. Statistics of the NYT08 and TRC2 newswire collections.

Quantity	NYT08	TRC2
Number of unique stories	102,853	1,800,370
Start date	01/01/08	01/01/08
End date	28/02/09	28/02/09
Content provided per story	Timestamp	Timestamp
	Article headline	Article headline
	Article URL	Article summary

the same period of time as the Blogs08 blog post corpus described previously (Section 6.1.2). The New York Times corpus, denoted NYT08, was used during the TREC 2009 task [136], while the Reuters corpus, denoted TRC2, was used during TREC 2010 [117]. For each of these news corpora, for a set of "topic days," stories published on those days were assessed in terms of their newsworthiness. Table 6.3 provides statistics for the NYT08 and TRC2 newswire collections.

6.4 Publication Venues

Work into the blogosphere has been accepted at various venues within the information retrieval community. Notable examples include:

- Advances in Research and Development in Information Retrieval (SIGIR) — is an annually held ACM conference, and represents the worldwide best information retrieval research.
- Conference in Information and Knowledge Management (CIKM) — is also an annually held ACM conference, and represents excellent research from the IR, database and knowledge management fields.
- European Conference in IR (ECIR) — is an information retrieval conference that is held annually at a European venue under the auspices of the British Computer Society.
- Web Search and Data Mining (WSDM) — is a young, annual ACM conference intended to be the publication venue for research in the areas of search and data mining.

- World Wide Web Conference (WWW) — is held annually, organised jointly by the World Wide Web Consortium (W3C) and the ACM. Among the various tracks, the Relevance and Ranking track contains prestigious information retrieval papers with a web search focus.
- Weblogging and Social Media (ICWSM) — is a maturing annual conference run under the auspices of the Association for the Advancement of Artificial Intelligence (AAAI). ICWSM grew out of the Workshop on the Weblogging Ecosystem (WWE) that last ran at WWW 2006. ICWSM accepts papers covering topics on data mining and search within social media, such as forums, blogs, and Twitter.
- Empirical Methods on Natural Language Processing (EMNLP) — is an annual natural language processing conference, which often accepts papers covering social media topics, including blogs.
- North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT) — is a biennial natural language processing conference in which papers on social media or sentiment analysis are often published.

7

Future Work in Blog and Microblog Search

We have surveyed current research and applications tackling various retrieval tasks within the blogosphere. In particular, three main tasks have been described, along with the most recent advances to effectively handle these tasks. In Section 3, we have described the retrieval of blog posts, as needed for context queries, along with related approaches for the identification of opinionated content within the blog posts. Section 4 addressed retrieving results for concept queries, which are usually answered by suggesting full blogs that the searcher would likely find interesting, read or subscribe to in their RSS reader. Section 5 described applications built upon searching and mining the blogosphere, such as identifying the top news stories, or analysing trends within the blogosphere. Finally, in Section 6, we have surveyed a number of resources produced by researchers over the past years to facilitate experimentation within a blogosphere context.

In this concluding section, we provide an overview of possible research directions in the field of blog search, with references to the material covered in the preceding sections. Furthermore, recently, interest in the community has shifted from the blogosphere at large into social networks [71], and social search [87], specifically microblogging.

As such, we provide an introduction to this emerging field and the new research directions that it is pursuing.

7.1 Blog Search

As demonstrated in Section 3, there is a huge body of work addressing opinion mining and retrieval within the blogosphere. This line of research by the community has been successful, and is rapidly maturing, assisted by the availability of common shared datasets with which to measure progress. An interesting area that is likely to grow is the identification and retrieval of a set of diverse opinions covering the query topics [48], combining opinion mining and search results diversification approaches such as [191]. Indeed, Living Knowledge[1] is a European FP7 project entirely dedicated to the investigation of the effect of diversity and time on opinions and bias.

The results in Section 4 show that while effective models exist for ranking blogs, the machine learning of models for different facets of blog search (as exemplified by [60, 63]) is still a hard problem. However, some more promising research has examined the temporal factors involved in searching blog posts [97, 98], as well as the identification of authoritative and influential bloggers [2, 58, 94, 108, 163], despite the relatively sparser link structure underlying the blogosphere [3]. On the other hand, none of these approaches has effectively made use of a key distinctive characteristic of the blogosphere when compared to most other Web content, namely, that blogs represent individuals. Indeed, effectively exploiting this view of the blogosphere as a social network for improving search systems is an open research direction [154].

As Section 5 clearly shows, blog-aided search is a recent and growing research area. The recent TREC 2009/2010 Blog track top stories identification task [136] sparked interest both from academia and industry (exemplified by the provision by Thomson Reuters of resources for tackling top stories detection) regarding the potential for the blogosphere to support new time-oriented and news-related search tasks. Accurate and timely trend-detection using the blogosphere continues to be an active

[1] http://livingknowledge-project.eu/.

research area [181, 193] with commercial interest from large brand monitoring companies, e.g., Nielsen Media Research.

7.2 Microblog Search

From the year 2009 onwards, there has been a shift in research effort from blogging trends and analysis to the new field of microblogs. This is not to say that the importance of blogs have diminished, indeed they remain a highly popular communication and information sharing medium [204]. For example, the popular blogging tool WordPress is used to make approximately 6.5 million posts each day.[2] Rather, researchers are now looking more to challenges that can be addressed by microblogs, e.g., real-time event detection [180, 189] or identification of recent Web pages of interest [49]. A microblog, as its name suggests, is a blog, but each post within that blog is of limited length. The idea behind microblogging is that by promoting short posts, real-time updates can be made with limited effort, possibly from outwith a typical desktop environment. For example, many microblogging platforms facilitate posting via text messages (SMS).

Twitter[3] is at the time of writing the largest dedicated English microblogging service. Twitter enables anyone with Internet access to sign-up and publicly post messages not exceeding 140 characters about any topic. It is highly popular, and currently has over 100 million active users generating 250 million tweets per day [59]. Other services that contain microblogging elements include Facebook[4] and Tumblr.[5] Another important platform is Sina Weibo, China's most popular microblogging platform, that is of growing interest to the multi-lingual information retrieval community [183].

Importantly, Twitter provides programmatic access to both its search service, and to a portion of its content stream as open and freely available APIs.[6] For this reason, much of the research into microblogging has focused on Twitter exclusively. For instance, conferences such

[2] http://en.wordpress.com/stats/posting/.
[3] http://www.twitter.com/.
[4] http://www.facebook.com/.
[5] http://www.tumblr.com.
[6] http://dev.twitter.com/.

as WSDM, ICWSM or EMNLP (see Section 6.4) are increasingly receiving submissions of papers using Twitter data. However, the terms of service for use of these APIs prohibit the re-distribution of tweet texts. This restriction has until recently slowed the creation of re-usable tweet test collections [198]. Twitter has some notable characteristics that influence how it is used. In particular, users can follow other users, creating a default time stream or "wall" containing all of the recent tweets by the people that the user follows in reverse-chronological order. This follower/followee relationship creates a rich social graph of users [110]. When posting, hashtags, i.e., words beginning with the character "#" are used to denote topics and concepts. These are used to link together many tweets about the same topic. Indeed, it has been reported that over 15% of tweets contain hashtags [51]. Similarly, mentions, i.e., user names prefixed with the "@" symbol are used to indicate replies or direct messages to the user in question. Additionally, Twitter allows a user to retweet another's tweets, i.e., post an exact copy of another user's tweet, normally with a reference to the source user [27]. This can be done in two manners. Firstly, as an automatic retweet, that just copies the post in question with Twitter itself logging and displaying the originating user. Alternatively, the user can retweet manually, whereby the user copies the tweet and adds RT:@USER to the front of it, optionally with a comment. Finally, when activated, Twitter automatically stores information regarding the geographical location of the person tweeting. Indeed, various uses for this information are already being explored, from earthquake detection [189] to location-based interaction analysis [188]. All of these peculiarities create an interesting environment for novel research investigation.

Tweet search also shares some of the blog search attributes examined in Section 2:

- Users can be looking for the most recent information about a real-world event (a context query), typified by reverse chronological rankings returned in response to queries. Indeed, the main challenge for participants of the recent TREC 2011 Microblog track real-time search task was to identify relevant tweets to display in reverse chronological

order. Correspondingly, the main challenge for the track organizers was to develop appropriate pooling and evaluation methodologies to estimate the performance of the participating systems at ranking the relevant tweets given the reverse chronological order of rankings.

- Alternatively, users can be looking for other users to "follow" that share their interests — indeed, both Twitter and Facebook deploy friend suggestions without the need for a query (called Who to Follow and People You May Know, respectively), based on the user's social network and content. This last need directly corresponds to concept queries.
- Finally, other tasks can be aided by the use of Twitter. For example, the detection and tracking of breaking news events is possible, as illustrated by Petrovic et al. [180].

However, search in a microblog setting also holds some notable differences. Teevan et al. [205] identified three types of information that users typically search for in a microblogging setting, namely: Timely information; Social information; and Topical information. Indeed, the temporal aspect of microblogging is particularly important, with users often searching for breaking news, real-time content and popular trends. Furthermore, for Twitter, queries have been shown to typically be even shorter than those observed for a Web search setting due to users searching with single term queries, e.g., hashtags or mentions [205].

Traditional *ad hoc* search on tweets has seen some investigation, but remains a challenging problem due to the short length of each tweet. In particular, Duan et al. [50] first examined how tweets can be ranked using machine learned approaches. They applied a learning to rank [128] approach using content relevance features, Twitter-specific features as well as account authority features for tweet ranking. They show that tweet length and the presence of a URL are important features for finding relevant tweets. Relatedly, prior to the development of the social network and microblogging platform Google+,[7] the Google search engine ranked tweets for display in its Web search results for

[7] https://plus.google.com/.

queries requiring recent content, e.g., breaking news queries [46]. The idea behind this was to serve fresh discussions from Twitter within the search results, when other forms of content about the topic, e.g., news articles, may not have yet been published. Google Fellow Amit Singhal gave an interview about the techniques employed for their Twitter search application [46]. He highlighted the importance of effectively using the author's user graph as an indicator of authority for finding high quality tweets. Indeed, how to best exploit the Twitter social graph is a new and important research area [34, 110]. Moreover, Singhal also revealed that Google considered hashtags to be a key indicator when combating spam tweets, another emerging research direction [227].

However, it remains an open question whether a traditional relevance ranking of tweets is the best way to satisfy users in a microblog setting [162]. For example, Twitter provides tweet rankings in reverse-chronological order, encapsulating the temporal nature of microblog search. In this case, there is a trade-off between returning fresh tweets that, while recent, may add little value, or returning older tweets that are more relevant but risk being seen as out-of-date. How this trade-off should be tackled to achieve effective search from a user perspective has yet to be fully explored. Other search tasks such as author/expert search have also been proposed in a microblog setting [162], but have not seen substantial investigation yet.

In 2011 the Text REtrieval Conference (TREC) ran a pilot Microblog track investigating *ad hoc* tweet search [172].[8] The aim of this task was to find the most relevant tweets for the user query in a real-time setting, i.e., to retrieve tweets on or before a point in time. However, in contrast to other *ad hoc* search tasks, the results returned were ranked in reverse chronological order. This ranking approach was to mimic search as seen on Twitter itself. The aim of the task was to find the most recent and relevant tweets for a user query.

To facilitate this track, the first legally redistributable Twitter test collection, named Tweets11 [145, 198], was developed through collaboration between TREC and Twitter [198]. Indeed, a unique feature of this test collection is the distribution mechanism, which facilitates the

[8] https://sites.google.com/site/microblogtrack/.

sharing of a standard tweets dataset while respecting Twitter's Terms and Conditions. In particular, the collection is distributed as a set of tweet unique identifiers ("tweet ids"), in conjunction with tools allowing the crawling of the data from the Twitter Web site. The necessary tools for downloading this test collection are available from the TREC Web site.[9]

For the 2011 *ad hoc* task, participant systems returned tweets for a point in time in reverse chronological order for a query. Systems were evaluated in terms of the number of relevant tweets they returned in the top 30 results. To rank tweets in this manner, a variety of approaches were proposed. For example, Amati et al. [10] proposed a new DFReeKLIM retrieval model from the divergence from randomness framework [7] that accounts for the very short nature of tweets. In contrast, McCreadie et al. [144] proposed a learning to rank framework for filtering a real-time stream of tweets of those that are low quality, irrelevant or not predominantly in English. The most effective approaches to the 2011 *ad hoc* task focused purely on relevance [172]. In particular, the approach by Metzler and Cai [149] combined learning to rank with pseudo-relevance feedback to find the 30 most relevant tweets to the user query and returned only those 30 tweets.

Overall, the field of Twitter search is still emerging, with both many promising lines of research and large volumes of rich data being made available. The TREC Microblog track is an ongoing effort, examining both real-time search in addition to introducing a new filtering task.[10] As with blogs, areas such as real-time news search [205], automatic news ranking and distillation tasks (*"find me twitterers that are interested in x"*) [77] are promising lines for research. Moreover, techniques to effectively exploit the rich social graph [110] of users may also lead to interesting insights for improving search, e.g., for identifying small groups of users expert in a particular topic [120] or for spam detection [218].

As this hot research area matures, and the Microblog track addresses more search tasks, we expect further publications to emerge

[9] http://trec.nist.gov/data/tweets/.
[10] https://sites.google.com/site/microblogtrack/.

in the near future in venues such as those described in Section 6.4. We expect these tasks to not only be about search directly, but also search aiding within a bigger context, such as at the boundary of sensor search and social search addressed by the recent SMART European FP7 project.[11] For instance, in this project, current events within a city can be detected using sensors (camera, audio sensors, traffic conditions), combined with social evidence (friends activities, geo-located tweets combined with sentiment analysis) to address tasks like *what interesting events are happening near a user* [199] or *which parts of the city are crowded right now?*

In conclusion, blog search is a venerable research area with many challenges yet to be solved, as discussed throughout this survey. On the other hand, microblog search is a growing research area with many new and exciting research directions to pursue [50, 110, 120, 183, 205, 218]. Search tasks using the blogosphere are well supported by publicly available test collections, many of which have been produced by TREC [131, 136]. Meanwhile, with the large volumes of microblog data being made available to the academic community from companies like Twitter and the development of new shared test collections such as Tweets11 [145, 198], there is much scope for further research into this field. Furthermore, as new platforms like Google+ are still emerging and existing ones evolve their business models, blog and microblog search are likely to remain hot topics in the information retrieval community and beyond for many years to come.

[11] http://www.smartfp7.eu/.

Acknowledgments

We would like to acknowledge the efforts of the participants in the TREC Blog track over the years 2006–2010. We are also thankful to the anonymous reviewers for their valuable feedback on this survey.

References

[1] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: divided they blog," in *Proceedings of the International Workshop on Link Discovery*, pp. 36–43, 2005.

[2] E. Adar, L. Zhang, L. Adamic, and R. Lukose, "Implicit structure and the dynamics of blogspace," in *Proceedings of the Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.

[3] N. Agarwal and H. Liu, "Blogosphere: research issues, tools, and applications," *SIGKDD Explorations Newsletter*, vol. 10, no. 1, pp. 18–31, 2008.

[4] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM 2008)*, pp. 207–218, 2008.

[5] N. F. Ali-Hasan and L. A. Adamic, "Expressing social relationships on the blog through links and comments," in *Proceedings of the International Conference on Weblogs and Social Media*, 2007.

[6] J. Allan, C. Wade, and A. Bolivar, "Retrieval and novelty detection at the sentence level," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR 2003)*, (New York, NY, USA), pp. 314–321, 2003.

[7] G. Amati, "Probability models for information retrieval based on divergence from randomness," University of Glasgow, 2003.

[8] 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 European Conference on IR Research on Advances in Information Retrieval (ECIR 2008)*, pp. 89–100, 2008.

[9] G. Amati, G. Amodeo, M. Bianchi, C. Gaibisso, and G. Gambosi, "FUB, IASI-CNR and University of Tor Vergata at TREC 2008 blog track," in *Proceedings of the Text REtrieval Conference*, (Gaithersburg, MD, USA), 2008.

[10] G. Amati, G. Amodeo, M. Bianchi, G. Marcone, C. Gaibisso, A. Celi, C. D. Nicola, and M. Flammini, "FUB, IASI-CNR, UNIVAQ at TREC 2011," in *Proceedings of the Text REtrieval Conference (TREC 2011)*, 2011.

[11] G. Amati, G. Amodeo, V. Capozio, C. Gaibisso, and G. Gambosi, "On performance of topical opinion retrieval," in *Proceedings of the international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2010)*, pp. 777–778, 2010.

[12] A. Andreevskaia, S. Bergler, and M. Urseanu, "All blogs are not made equal: exploring genre differences in sentiment tagging of blogs," in *Proceedings of the AAAI International Conference on Weblogs and Social Media (ICWSM 2007)*, 2007.

[13] 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.

[14] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proceedings of the Conference on Language Resources and Evaluation (LREC 2010)*, 2010.

[15] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 1999.

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

[17] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet Project," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics (ACL 1998)*, pp. 86–90, 1998.

[18] T. Ballmer and W. Brennenstuhl, *Speech Act Classification: A Study of the Lexical Analysis of English Speech Activity Verbs*. Springer-Verlag, 1981.

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

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

[21] N. Bansal and N. Koudas, "BlogScope: spatio-temporal analysis of the blogosphere," in *Proceedings of the International Conference on World Wide Web*, ACM, pp. 1269–1270, 2007.

[22] R. Berkman, *The Art of Strategic Listening: Finding Market Intelligence Through Blogs and Other Social Media*. Paramount Market Publishing, 2008.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[24] R. Blood, *The Weblog Handbook: Practical Advice on Creating and Maintaining Your Blog*. Perseus Publishing, 2002.

[25] R. Blood, "How blogging software reshapes the online community," *Communications of the ACM*, vol. 47, no. 12, pp. 53–55, 2004.

[26] A. A. Bolourian, Y. Moshfeghi, and C. J. van Rijsbergen, "Quantification of topic propagation using percolation theory: A study of the icwsm network," in *Proceedings of the Annual Conference on Weblogs and Social Media (ICWSM 2009)*, 2009.

[27] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," *Hawaii International Conference on System Sciences*, pp. 1–10, 2010.

[28] A. Broder, "A taxonomy of web search," *SIGIR Forum*, vol. 36, no. 2, pp. 3–10, 2002.

[29] K. Burton, A. Java, and I. Soboroff, "The icwsm 2009 spinn3r dataset," in *Proceedings of the Annual Conference on Weblogs and Social Media (ICWSM 2009)*, 2009.

[30] K. Burton, N. Kasch, and I. Soboroff, "The icwsm 2011 spinn3r dataset," in *Proceedings of the Annual Conference on Weblogs and Social Media (ICWSM 2011)*, 2011.

[31] 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*, vol. 41, no. 5, pp. 1141–1161, 2005.

[32] J. Callan, "Distributed information retrieval," in *Advances in Information Retrieval*, (W. B. Croft, ed.), Kluwer Academic Publishers, pp. 127–150, 2000.

[33] C. Castillo and B. D. Davison, "Adversarial web search," *Foundations and Trends in Information Retrieval*, vol. 4, no. 5, pp. 377–486, 2010.

[34] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 10–17, 2010.

[35] M. Cha, J. Perez, and H. Haddadi, "Flash floods and ripples: The spread of media content through the blogosphere," in *Proceedings of the Annual Conference on Weblogs and Social Media (ICWSM 2009)*, 2009.

[36] D. Chakrabarti, R. Kumar, and K. Punera, "Page-level template detection via isotonic smoothing," in *Proceedings of the International Conference on World Wide Web, (WWW '07)*, pp. 61–70, 2007.

[37] J. M. Chenlo and D. E. Losada, "Combining document and sentence scores for blog topic retrieval," in *Proceedings of the Spanish Conference on Information Retrieval (CERI 2010)*, 2010.

[38] J. M. Chenlo and D. E. Losada, "Effective and efficient polarity estimation in blogs based on sentence-level evidence," in *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM 2011)*, 2011.

[39] Y. Chi, B. L. Tseng, and J. Tatemura, "Eigen-trend: trend analysis in the blogosphere based on singular value decompositions," in *Proceedings of the*

ACM International Conference On Information and Knowledge Management, ACM, pp. 68–77, 2006.

[40] D. Chinavle, P. Kolari, T. Oates, and T. Finin, "Ensembles in adversarial classification for spam.," in *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM 2009)*, (D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, and J. J. Lin, eds.), ACM, pp. 2015–2018, 2009.

[41] J. Cho and A. Tomkins, "Social media and search," *IEEE Internet Computing*, vol. 11, no. 6, pp. 13–15, 2007.

[42] C. Cleverdon, "The cranfield tests on index language devices," *Aslib Proceedings*, vol. 19, no. 6, pp. 173–194, 1967.

[43] D. Cohn and T. Hofmann, "The missing link: a probabilistic model of document content and hypertext connectivity," in *Neural Information Processing Systems 13*, pp. 430–436, 2000.

[44] N. Craswell, A. P. de Vries, and I. Soboroff, "Overview of the TREC-2005 enterprise track," in *Proceedings of the Text REtrieval Conference (TREC-2005)*, *vol. 500–266 of NIST Special Publication*, 2006.

[45] N. Craswell and M. Szummer, "Random walks on the click graph," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 239–246, 2007.

[46] C. Crum, "Google reveals factors for ranking tweets," 2010. http://www.webpronews.com/google-reveals-factors-for-ranking-tweets-2010-01, accessed on 29/09/2011.

[47] danah michele boyd, "Taken out of context — American teen sociality in networked publics," PhD Thesis, University of California, Berkeley, 2008.

[48] G. Demartini, "ARES: a retrieval engine based on sentiments sentiment-based search result annotation and diversification," in *Proceedings of the European Conference on Advances in Information Retrieval (ECIR 2011)*, pp. 772–775, 2011.

[49] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha, "Time is of the essence: improving recency ranking using twitter data," in *Proceedings of the International Conference on World Wide Web*, ACM, pp. 331–340, 2010.

[50] Y. Duan, L. Jiang, T. Qin, M. Zhou, and H. Shum, "An empirical study on learning to rank of tweets," in *Proceedings of the International Conference on Computational Linguistics*, pp. 295–303, 2010.

[51] M. Efron, "Information search and retrieval in microblogs," *Journal of the American Society for Information Science and Technology*, 2011.

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

[53] B. Ernsting, W. Weerkamp, and M. de Rijke, "Language modeling approaches to blog post and feed finding," in *Proceedings of the Text REtrieval Conference*, 2007.

[54] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed-membership models of scientific publications," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5220–5227, 2004.

[55] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," in *Proceedings of the Conference on Language Resources and Evaluation (LREC 2006*, pp. 417–422, 2006.

[56] S. Evert, "A lightweight and efficient tool for cleaning web pages," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC'08)*, (N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odjikand, S. Piperidis, and D. Tapias, eds.), (Marrakech, Morocco), May 2008. http://www.lrec-conf.org/proceedings/lrec2008/.

[57] L. Franco and H. Kawai, "News detection in the blogosphere: two approaches based on structure and content analysis," in *Proceedings of the Data Challenge Workshop at ICWSM*, 2010.

[58] K. Fujimura, T. Inoue, and T. Inoue, "The EigenRumor algorithm for ranking blogs," in *Proceedings of the Annual Workshop on the Weblogging Ecosystem (WWE 2005)*, 2005.

[59] L. Gannes, "Twitter dumps on google for pushing google+ in search," 2012. http://allthingsd.com/20120110/twitter-dumps-on-google-for-pushing-google-plus-in-search/, accessed on 12/01/2012.

[60] D. Gao, R. Zhang, W. Li, Y. K. Lau, and K. F. Wong, "Learning features through feedback for blog distillation," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information (SIGIR 2011)*, pp. 1085–1086, 2011.

[61] S. Gerani, M. Carman, and F. Crestani, "Proximity based opinion retrieval," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*, 2010.

[62] S. Gerani, M. J. Carman, and F. Crestani, "Investigating learning approaches for blog post opinion retrieval," in *Proceedings of the European Conference on IR Research on Advances in Information Retrieval (ECIR 2009)*, pp. 313–324, 2009.

[63] S. Gerani, M. Keikha, M. Carman, and F. Crestani, "Personal blog retrieval using opinion features," in *Proceedings of the European Conference on Advances in Information Retrieval (ECIR 2011)*, pp. 747–750, 2011.

[64] S. Gerani, M. Keikha, and F. Crestani, "Aggregating multiple opinion evidence in proximity-based opinion retrieval," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information (SIGIR 2011)*, pp. 1199–1200, 2011.

[65] A. J. Gill, S. Nowson, and J. Oberlander, "What are they blogging about? personality, topic and motivation in blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2009)*, 2009.

[66] K. E. Gill, "How can we measure the influence of the blogosphere?," in *Proceedings of the Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.

[67] D. Gillmor, *We the Media: Grassroots Journalism by the People, for the People. O'Reilly Series*, O'Reilly, 2006.

[68] N. S. Glance, M. Hurst, and T. Tomokiyo, "BlogPulse: automated trend discovery for weblogs," in *Proceedings of the Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.

[69] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, pp. 219–222, 2007.

[70] A. Gordon and R. Swanson, "Identifying personal stories in millions of weblog entries," in *Proceedings of the Annual Conference on Weblogs and Social Media (ICWSM 2009)*, 2009.

[71] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *Proceedings of the ACM International Conference on Web Search and Data Mining, (WSDM '10)*, (New York, NY, USA), pp. 241–250, 2010.

[72] M. L. Gregory, D. Payne, D. McColgin, N. Cramer, and D. Love, "Visual analysis of weblog content," in *Proceedings of the AAAI International Conference on Weblogs and Social Media (ICWSM 2007)*, 2007.

[73] D. Gruhl, L. Chavet, D. Gibson, J. Meyer, P. Pattanayak, A. Tomkins, and J. Zien, "How to build a WebFountain: an architecture for very large-scale text analytics," *IBM Systems Journal*, vol. 43, no. 1, pp. 64–77, 2004.

[74] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proceedings of the International Conference on World Wide Web, (WWW 2004)*, pp. 491–501, 2004.

[75] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of trust and distrust," in *Proceedings of the International Conference on World Wide Web (WWW 2004)*, pp. 403–412, 2004.

[76] D. Hannah, C. Macdonald, J. Peng, B. He, and I. Ounis, "University of Glasgow at TREC 2007: Experiments in blog and enterprise tracks with terrier," in *Proceedings of the Text REtrieval Conference*, 2007.

[77] J. Hannon, M. Bennett, and B. Smyth, "Recommending twitter users to follow using content and collaborative filtering approaches," in *Proceedings of the ACM Conference on Recommender Systems*, ACM, pp. 199–206, 2010.

[78] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (ACL 1997)*, Association for Computational Linguistics, pp. 174–181, 1997.

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

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

[81] M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, and K.-P. Yee, "Finding the flow in web site search," *Communications ACM*, vol. 45, pp. 42–49, September 2002.

[82] M. Hearst, M. Hurst, and S. Dumais, "What should blog search look like?," in *Proceedings of the International Workshop on Search in Social Media (SSM-2008)*, 2008.

[83] M. A. Hearst, *Search User Interfaces*. Cambridge University Press, 2009.
[84] M. A. Hearst and S. T. Dumais, "Blogging together: An examination of group blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2009)*, 2009.
[85] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of Uncertainty in Artificial Intelligence (UAI 1999)*, pp. 289–296, 1999.
[86] S. Holtz, J. Havens, and L. Johnson, *Tactical Transparency: How Leaders can Leverage Social Media to Maximize Value and Build their Brand*. Vol. 6, Jossey-Bass Inc Pub, 2009.
[87] D. Horowitz and S. D. Kamvar, "The anatomy of a large-scale social search engine," in *Proceedings of the International Conference on World Wide Web, (WWW '10)*, (New York, NY, USA), pp. 431–440, 2010.
[88] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, pp. 168–177, 2004.
[89] X. Huang and W. B. Croft, "A unified relevance model for opinion retrieval," in *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM 2009)*, ACM, pp. 947–956, 2009. ISBN 978-1-60558-512-3. doi: http://doi.acm.org/10.1145/1645953.1646075.
[90] 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)*, University of Maryland, Baltimore County, 2007.
[91] L. Jia, C. Yu, and W. Meng, "The effect of negation on sentiment analysis and retrieval effectiveness," in *Proceeding of the ACM Conference on Information and Knowledge Management (CIKM 2009)*, pp. 1827–1830, 2009.
[92] L. Jia, C. T. Yu, and W. Zhang, "UIC at TREC 2008 blog track," in *Proceedings of the Text REtrieval Conference*, 2008.
[93] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the International Conference on Web Search and Web Data Mining, (WSDM '08)*, (New York, NY, USA), pp. 219–230, 2008.
[94] A. Kale, A. Karandikar, P. Kolari, A. Java, and A. Joshi, "Modeling trust and influence in the blogosphere using link polarity," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, 2007.
[95] M. Keikha and F. Crestani, "Effectiveness of aggregation methods in blog distillation," in *Proceedings of the International Conference on Flexible Query Answering Systems (FQAS 2009)*, pp. 157–167, 2009.
[96] M. Keikha and F. Crestani, "Linguistic aggregation methods in blog retrieval," *Information Processing and Management*, 2011.
[97] M. Keikha, S. Gerani, and F. Crestani, "Relevance stability in blog retrieval," in *Proceedings of the 2011 ACM Symposium on Applied Computing (SAC 2011)*, pp. 1119–1123, 2011.
[98] M. Keikha, S. Gerani, and F. Crestani, "Temper: A temporal relevance feedback method," in *Proceedings of the European Conference on Advances in Information Retrieval (ECIR 2011)*, pp. 436–447, 2011.

[99] A. King, "The evolution of RSS," April 2004. http://www.webreference.com/authoring/languages/xml/rss/1/, last accessed 14/09/2006.

[100] J. Kleinberg, "Hubs, authorities, and communities," *ACM Computing Surveys (CSUR)*, vol. 31, no. 4es, p. 5, 1999.

[101] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," in *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 1998)*, pp. 668–677, 1998.

[102] C. Kohlschütter, P. Fankhauser, and W. Nejdl, "Boilerplate detection using shallow text features," in *Proceedings of the ACM International Conference on Web Search and Data Mining, (WSDM '10)*, (New York, NY, USA), pp. 441–450, 2010.

[103] P. Kolari, T. Finin, and A. Joshi, "SVMs for the blogosphere: Blog identification and splog detection," in *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[104] P. Kolari, A. Java, and T. Finin, "Characterizing the splogosphere," in *Proceedings of the Annual Workshop on the Weblogging Ecosystem (WWE 2006)*, 2006.

[105] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi, "Detecting spam blogs: a machine learning approach," in *Proceedings of the National Conference on Artificial Intelligence (AAAI 2006)*, 2006.

[106] A. C. König, M. Gamon, and Q. Wu, "Click-through prediction for news queries," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, ACM, pp. 347–354, 2009.

[107] A. Kritikopoulos and M. Sideri, "The compass filter: search engine result personalization using web communities (itwp 2003)," in *Intelligent Techniques for Web Personalization Workshop (ITWP 2003) at IJCAI*, pp. 229–240, 2003.

[108] A. Kritikopoulos, M. Sideri, and I. Varlamis, "BlogRank: ranking weblogs based on connectivity and similarity features," in *Proceedings of International Workshop on Advanced Architectures and Algorithms for Internet DElivery and Applications (AAA-IDEA 2006)*, ACM, 2006.

[109] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "On the bursty evolution of blogspace," in *Proceedings of the International Conference on World Wide Web (WWW 2003)*, pp. 568–576, 2003.

[110] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media," in *WWW '10: Proceedings of the International World Wide Web Conference*, (New York, NY, USA), 2010.

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

[112] R. Lau, R. Rosenfeld, and S. Roukos, "Trigger-based language models: a maximum entropy approach," in *Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing: Speech Processing — Volume II (ICASSP 1993)*, pp. 45–48, 1993.

[113] V. Lavrenko and W. B. Croft, "Relevance based language models," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pp. 120–127, 2001.

[114] Y. Lee, H.-Y. Jung, W. Song, and J.-H. Lee, "Mining the blogosphere for top news stories identification," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*, 2010.

[115] Y. Lee, S.-H. Na, J. Kim, S.-H. Nam, H.-Y. Jung, and J.-H. Lee, "KLE at TREC 2008 Blog track: blog post and feed retrieval," in *Proceedings of the Text REtrieval Conference*, 2008.

[116] 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 ACM conference on Information and Knowledge Management (CIKM 2009)*, pp. 1971–1974, 2009.

[117] J. L. Leidner, "Thomson reuters releases trc2 news corpus through nist," 2010. http://jochenleidner.posterous.com/thomson-reuters-releases-research-collection, accessed on 16/01/2011.

[118] B. Levin, *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, 1993.

[119] B. Li, F. Liu, and Y. Liu, "UTDallas at TREC 2008 blog track," in *Proceedings of the Text REtrieval Conference*, 2008.

[120] Q. Liao, C. Wagner, P. Pirolli, and W. Fu, "Understanding experts' and novices' expertise judgment of twitter users," in *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems*, ACM, pp. 2461–2464, 2012.

[121] Y.-F. Lin, J.-H. Wang, L.-C. Lai, and H.-Y. Kao, "Top stories identification from blog to news in trec 2010 blog track," in *Proceedings of the Text REtrieval Conference*, 2010.

[122] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. Tseng, "Discovery of blog communities based on mutual awareness," in *Proceedings of the Annual Workshop on the Weblogging Ecosystem (WWE 2006)*, 2006.

[123] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. L. Tseng, "Splog detection using self-similarity analysis on blog temporal dynamics," in *Proceedings of the International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2007)*, pp. 1–8, 2007.

[124] C. Lioma, C. Macdonald, V. Plachouras, J. Peng, B. He, and I. Ounis, "University of Glasgow at TREC 2006: Experiments in terabyte and enterprise tracks with terrier," in *Proceedings of the Text REtrieval Conference (TREC 2006)*, 2006.

[125] F. Liu, B. Li, and Y. Liu, "Finding opinionated blogs using statistical classifiers and lexical features," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2009)*, 2009.

[126] J. Liu, L. Birnbaum, and B. Pardo, "Spectrum: Retrieving different points of view from the blogosphere," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2009)*, 2009.

[127] S. Liu, F. Liu, C. Yu, and W. Meng, "An effective approach to document retrieval via utilizing wordnet and recognizing phrases," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pp. 266–272, 2004.

[128] T.-Y. Liu, "Learning to rank for information retrieval," *Foundations and Trends in Information Retrieval*, vol. 3, pp. 225–331, 2009.

[129] C. Macdonald, B. He, I. Ounis, and I. Soboroff, "Limits of opinion-finding baseline systems," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 747–748, 2008.

[130] C. Macdonald, B. He, V. Plachouras, and I. Ounis, "University of glasgow at trec 2005: Experiments in terabyte and enterprise tracks with terrier," in *Proceedings of the Text REtrieval Conference (TREC 2005)*, 2005.

[131] 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.

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

[133] C. Macdonald and I. Ounis, "Key blog distillation: ranking aggregates," in *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM 2008)*, ACM, pp. 1043–1052, 2008.

[134] C. Macdonald, I. Ounis, and I. Soboroff, "Overview of the TREC-2007 blog track," in *Proceedings of the Text REtrieval Conference*, 2007.

[135] C. Macdonald, I. Ounis, and I. Soboroff, "Is spam an issue for opinionated blog post search?," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, ACM, pp. 710–711, 2009.

[136] C. Macdonald, I. Ounis, and I. Soboroff, "Overview of the TREC 2009 Blog track," in *Proceedings of the Text REtrieval Conference*, 2009.

[137] C. Macdonald and I. Ounis, "Searching for expertise: Experiments with the Voting Model," *Computer Journal: Special Focus on Profiling Expertise and Behaviour*, vol. 52, no. 7, pp. 729–748, 2009.

[138] I. MacKinnon and O. Vechtomova, "Improving complex interactive question answering with wikipedia anchor text," in *Proceedings of the European Conference on IR Research on Advances in Information Retrieval (ECIR 2008)*, pp. 438–445, 2008.

[139] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[140] D. Matheson, "Weblogs and the epistemology of the news: Some trends in online journalism.," *New Media and Society*, vol. 6, no. 4, pp. 443–468, 2004.

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

[142] R. McCreadie, C. Macdonald, and I. Ounis, "A learned approach for ranking news in real-time using the blogosphere," in *Proceedings of the International Symposium on String Processing and Information Retrieval (SPIRE 2011)*, 2011.

[143] R. McCreadie, C. Macdonald, I. Ounis, J. Peng, and R. Santos, "University of glasgow at TREC 2009: Experiments with terrier," in *Proceedings of the Text REtrieval Conference*, 2009.

[144] R. McCreadie, C. Macdonald, R. Santos, and I. Ounis, "University of glasgow at trec 2011: Experiments with terrier in crowdsourcing, microblog, and web tracks," in *Proceedings of the Text REtrieval Conference (TREC 2011)*, 2011.

[145] R. McCreadie, I. Soboroff, J. Lin, C. Macdonald, I. Ounis, and D. McCullough, "On building a reusable twitter corpus," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012)*, 2012.

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

[147] Y. Mejova, V. H. Thuc, S. Foster, C. Harris, B. Arens, and P. Srinivasan, "Trec blog and trec chem: A view from the corn fields," in *Proceedings of the Text REtrieval Conference*, 2009.

[148] F. Mesquita, Y. Merhav, and D. Barbosa, "Extracting information networks from the blogosphere: state-of-the-art and challenges," in *Proceedings of the Data Challenge Workshop at ICWSM*, 2010.

[149] D. Metzler and C. Cai, "Usc/isi at trec 2011: Microblog track," in *Proceedings of the Text REtrieval Conference (TREC 2011)*, 2011.

[150] D. Metzler and W. B. Croft, "Combining the language model and inference network approaches to retrieval," *Information and Processing Management*, vol. 40, no. 5, pp. 735–750, 2004.

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

[152] M. Michelson and S. A. Macskassy, "What blogs tell us about websites: a demographics study," in *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM 2011)*, pp. 365–374, 2011.

[153] G. A. Miller, "Wordnet: a lexical database for english," *Communications ACM*, vol. 38, pp. 39–41, November 1995.

[154] G. Mishne, "Information access challenges in the blogspace," in *International Workshop on Intelligent Information Access*, 2006.

[155] G. Mishne, D. Carmel, and R. Lempel, "Blocking blog spam with language model disagreement," in *Proceedings of the International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005)*, pp. 1–6, 2005.

[156] G. Mishne and M. de Rijke, "A study of blog search," in *Proceedings of the European Conference on Information Retrieval (ECIR 2006)*, Springer, pp. 289–301, 2006.

[157] G. Mishne and N. Glance, "Leave a reply: an analysis of weblog comments," in *Proceedings of the Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.

[158] M. Montague and J. A. Aslam, "Condorcet fusion for improved retrieval," in *Proceedings of the International Conference on Information and Knowledge Management (CIKM 2002)*, (New York, NY, USA), pp. 538–548, 2002.

[159] S.-H. Na, I.-S. Kang, Y. Lee, and J.-H. Lee, "Applying completely-arbitrary passage for pseudo-relevance feedback in language modeling approach," in *Proceedings of the Asia Infomation Retrieval Symposium*, pp. 626–631, 2008.

[160] S.-H. Na, I.-S. Kang, Y. Lee, and J.-H. Lee, "Completely-arbitrary passage retrieval in language modeling approach," in *Proceedings of the Asia Infomation Retrieval Symposium*, pp. 22–33, 2008.

[161] 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 European Conference on IR Research on Advances in Information Retrieval (ECIR 2009)*, Springer-Verlag, pp. 734–738, 2009.

[162] R. Nagmoti, A. Teredesai, and M. D. Cock, "Ranking approaches for microblog search," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1, pp. 153–157, 2010.

[163] R. Nallapati and W. W. Cohen, "Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2008)*, 2008.

[164] S.-H. Nam, S.-H. Na, Y. Lee, and J.-H. Lee, "DiffPost: filtering non-relevant content based on content difference between two consecutive blog posts," in *Proceedings of the European Conference on Information Retrieval*, Springer, pp. 791–795, 2009.

[165] B. A. Nardi, D. J. Schiano, M. Gumbrecht, and L. Swartz, "Why we blog," *Communications of the ACM*, vol. 47, no. 12, pp. 41–46, 2004.

[166] M. Nottingham and R. Sayre, "The atom syndication format," Technical Report, The Internet Society, December 2005.

[167] S. Nowson and J. Oberlander, "Identifying more bloggers: towards large scale personality classification of personal weblogs," in *Proceedings of the AAAI International Conference on Weblogs and Social Media (ICWSM 2007)*, 2007.

[168] J. Oberlander and S. Nowson, "Whose thumb is it anyway?: classifying author personality from weblog text," in *Proceedings of the COLING/ACL on Main Conference Poster Sessions, (COLING-ACL '06)*, (Stroudsburg, PA, USA), Association for Computational Linguistics, pp. 627–634, 2006.

[169] N. A. of America (NAA), "Newspaper Web sites attract more than 70 million visitors in June; over one-third of all Internet users visit newspaper Web sites," 2010. http://www.naa.org/PressCenter/SearchPressReleases/2009/NEWSPAPER-WEB-SITES-ATTRACT-MORE-THAN-70-MILLION-VISITORS.aspx, accessed on 25/01/2010.

[170] M. Oka, H. Abe, and K. Kato, "Extracting topics from weblogs through frequency segments," in *Proceedings of the Annual Workshop on the Weblogging Ecosystem (WWE 2006)*, 2006.

[171] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff, "Overview of the TREC-2006 blog track," in *Proceedings of the Text REtrieval Conference*, 2006.

[172] I. Ounis, C. Macdonald, J. Lin, and I. Soboroff, "Overview of the TREC-2011 microblog track," in *Proceedings of the Text REtrieval Conference*, 2011.

[173] I. Ounis, C. Macdonald, and I. Soboroff, "Overview of the TREC-2008 blog track," in *Procedings of the Text REtrieval Confernece*, 2008.

[174] I. Ounis, C. Macdonald, and I. Soboroff, "Overview of the TREC 2010 blog track," in *Proceedings of the Text REtrieval Conference*, 2010.

[175] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: bringing order to the web," Technical Report 1999-66, Stanford InfoLab, 1999.

[176] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2010.

[177] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in *Proceedings of the International Conference on World Wide Web (WWW 2010)*, 2010.

[178] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[179] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pp. 79–86, 2002.

[180] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, *(HLT '10)*, (Stroudsburg, PA, USA), pp. 181–189, 2010.

[181] F. Pimenta, D. Obradovic, R. Schirru, S. Baumann, and A. Dengel, "Automatic sentiment monitoring of specific topics in the blogosphere," in *Workshop on Dynamic Networks and Knowledge Discovery (DyNaK 2010)*, 2010.

[182] J. C. Platt, *Probabilities for SV Machines*, pp. 61–74. MIT Press, 2000.

[183] Y. Qu, C. Huang, P. Zhang, and J. Zhang, "Microblogging after a major disaster in china: A case study of the 2010 yushu earthquake," in *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, ACM, pp. 25–34, 2011.

[184] J. Rettberg, *Blogging. Digital media and society series*, Polity Press, 2008.

[185] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proceedings of the Text REtrieval Conference (TREC 3)*, 1994.

[186] J. J. Rocchio, "Relevance feedback in information retrieval," in *The SMART Retrieval System: Experiments in Automatic Document Processing*, (G. Salton, ed.), Prentice Hall, pp. 313–323, 1971.

[187] A. Rosenbloom, "The blogosphere," *Communications of the ACM*, vol. 47, no. 12, pp. 30–33, 2004.

[188] A. Sadilek, H. Kautz, and J. Bigham, "Finding your friends and following them to where you are," in *Proceedings of the ACM International Conference on Web Search and Data Mining*, ACM, pp. 723–732, 2012.

[189] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *WWW '10: Proceedings of the International Conference on World Wide Web*, (New York, NY, USA), ACM, 2010.

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

[191] R. L. T. Santos, C. Macdonald, and I. Ounis, "Exploiting query reformulations for web search result diversification," in *Proceedings of the International Conference on World Wide Web (WWW 2010)*, pp. 881–890, 2010.

[192] 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)*, 2009.

[193] R. Schirru, D. Obradović, S. Baumann, and P. Wortmann, "Domain-specific identification of topics and trends in the blogosphere," in *Proceedings of the Industrial Conference on Advances in Data Mining (ICDM 2010)*, Springer-Verlag, pp. 490–504, 2010.

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

[195] J. Seo and W. B. Croft, "Blog site search using resource selection," in *Proceeding of the ACM Conference on Information and Knowledge Management (CIKM 2008)*, ACM, pp. 1053–1062, 2008.

[196] X. Shi, B. Tseng, and L. Adamic, "Looking at the blogosphere topology through different lenses," in *Proceedings of the AAAI International Conference on Weblogs and Social Media (ICWSM 2007)*, 2007.

[197] J. Sobel, "State of the Blogosphere," October 2010. http://technorati.com/blogging/article/state-of-the- blogosphere-2010-introduction.

[198] I. Soboroff, D. McCullough, J. Lin, C. Macdonald, I. Ounis, and R. McCreadie, "Evaluating real-time search over tweets," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2012)*, 2012.

[199] J. Soldatos, M. Draief, C. Macdonald, and I. Ounis, "Multimedia search over integrated social and sensor networks," in *Proceedings of the International Conference Companion on World Wide Web, (WWW '12) Companion*, (New York, NY, USA), pp. 283–286, 2012.

[200] S. Sood and L. Vasserman, "Sentisearch: Exploring mood on the web," in *Proceedings of the Annual Conference on Weblogs and Social Media (ICWSM 2009)*, 2009.

[201] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie, *The General Inquirer: A Computer Approach to Content Analysis.* MIT Press, 1966.

[202] J. Surowiecki, *The Wisdom of Crowds.* Doubleday, 2004.

[203] Syomos, "Inside blog demographics," June 2010. http://www.sysomos.com/reports/bloggers/.

[204] Technorati, "State of the Blogosphere 2011: Introduction and Methodology," 2011. http://technorati.com/social-media/article/state-of-the-blogosphere-2011-introduction/, accessed on 23/04/2011.

[205] J. Teevan, D. Ramage, and M. Morris, "# twittersearch: a comparison of microblog search and web search," in *Proceedings of the ACM International Conference on Web Search and Data Mining*, ACM, pp. 35–44, 2011.

[206] M. Thelwall, "Bloggers during the London attacks: Top information sources and topics," in *Proceedings of the International Workshop on the Weblogging Ecosystem*, 2006. http://www.blogpulse.com/www2006-workshop/papers/blogs-during-london-attacks.pdf.

[207] M. Thelwall, "Blog searching: The first general-purpose source of retrospective public opinion in the social sciences?," *Online Information Review*, vol. 31, no. 3, pp. 277–289, 2007.

[208] M. Thelwall and L. Hasler, "Blog search engines," *Online Information Review*, vol. 31, no. 4, pp. 467–479, 2007.

[209] V. H. Thuc, Y. Mejova, C. Harris, and P. Srinivasan, "Event intensity tracking in weblog collections," in *Proceedings of the Annual Conference on Weblogs and Social Media (ICWSM 2009)*, 2009.

[210] J. W. Treem and K. Y. Thomas, "What makes a blog a blog? exploring user conceptualizations of an old "new" online medium," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2010)*, 2010.

[211] P. Turney, M. Littman, R. Schirru, S. Baumann, and A. Dengel, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Transactions on Information Systems (TOIS)*, vol. 21, no. 4, 2003.

[212] P. D. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," *CoRR*, vol. cs.LG/0212032, 2002.

[213] O. Vechtomova, "University of waterloo at TREC 2008 blog track," in *Proceedings of the Text REtrieval Conference*, 2008.

[214] O. Vechtomova, "Facet-based opinion retrieval from blogs," *Information Processing and Management*, vol. 46, no. 1, pp. 71–88, 2010.

[215] K. Vieira, A. S. da Silva, N. Pinto, E. S. de Moura, J. a. M. B. Cavalcanti, and J. Freire, "A fast and robust method for web page template detection and removal," in *Proceedings of the ACM International Conference on Information and Knowledge Management, (CIKM '06)*, (New York, NY, USA), ACM, pp. 258–267, 2006.

[216] E. M. Voorhees, "Trec: Continuing information retrieval's tradition of experimentation," *Communications of the ACM*, vol. 50, no. 11, pp. 51–54, 2007.

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

[218] A. Wang, "Don't follow me: Spam detection in twitter," in *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, IEEE, pp. 1–10, 2010.

[219] W. Weerkamp, K. Balog, and M. de Rijke, "A two-stage model for blog feed search," in *Proceeding of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*, pp. 877–878, 2010.

[220] W. Weerkamp and M. de Rijke, "External query expansion in the blogosphere," in *Proceedings of the Text REtrieval Conference*, 2008.

[221] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2–3, pp. 165–210, 2005.

[222] 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*, pp. 34–35, 2005.

[223] X. Xu, Y. Liu, H. Xu, X. Yu, Z. Peng, X. Cheng, L. Xiao, and S. Nie, "ICTNET at Blog track TREC 2010," in *Proceedings of the Text REtrieval Conference*, 2010.

[224] R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decisionmaking," *IEEE Transactions Systems Man and Cybernetics*, vol. 18, pp. 183–190, 1988.

[225] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the International Conference on Machine Learning (ICML 1997)*, pp. 412–420, 1997.

[226] T. Yano and N. A. Smith, "What's worthy of comment? content and comment volume in political blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2010)*, 2010.

[227] S. Yardi, D. Romero, G. Schoenebeck, *et al.*, "Detecting spam in a twitter network," *First Monday*, vol. 15, no. 1, 2009.

[228] Y. Yue, T. Finley, F. Radlinski, and T. Joachims, "A support vector method for optimizing average precision," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 271–278, 2007.

[229] L. Zadeh, *A Computational Approach to Fuzzy Quantifiers in Natural Languages. Memorandum*, University of California, Berkeley, 1982.

[230] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proceedings of the International Conference on Information and Knowledge Management*, ACM, pp. 403–410, 2001.

[231] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Transactions Information Systems*, vol. 22, pp. 179–214, 2004.

[232] M. Zhang and X. Ye, "A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, pp. 411–418, 2008.

[233] Q. Zhang, B. Wang, L. Wu, and X. Huang, "Fdu at trec 2007: opinion retrieval of blog track," in *Proceedings of the Text REtrieval Conference (TREC 2007)*, 2007.

[234] W. Zhang, L. Jia, C. Yu, and W. Meng, "Improve the effectiveness of the opinion retrieval and opinion polarity classification," in *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM 2008)*, ACM, pp. 1415–1416, 2008.

[235] W. Zhang, S. Liu, C. Yu, C. Sun, F. Liu, and W. Meng, "Recognition and classification of noun phrases in queries for effective retrieval," in *Proceedings of the ACM Conference on Conference on Information and Knowledge Management (CIKM 2007)*, pp. 711–720, 2007.

[236] W. Zhang and C. Yu, "UIC at TREC 2006 blog track," in *Proceedings of the Text REtrieval Conference (TREC 2006)*, 2006.

[237] W. Zhang and C. Yu, "UIC at TREC 2007 blog track," in *Proceedings of the Text REtrieval Conference (TREC 2007)*, 2007.

[238] W. Zhang, C. Yu, and W. Meng, "Opinion retrieval from blogs," in *Proceedings of the ACM Conference on Conference on Information and Knowledge Management (CIKM 2007)*, ACM, pp. 831–840, 2007.

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

[240] K. Zhao, A. Kumar, M. Spaziani, and J. Yen, "Who blogs what: understanding behavior, impact and types of bloggers," in *Proceedings of the Annual Workshop on Information Technologies and Systems (WITS 2010)*, pp. 176–181, 2010.

[241] C.-N. Ziegler and M. Skubacz, "Toward automated reputation and brand monitoring on the web," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006)*, pp. 1066–1072, 2006.