Information Retrieval on the Blogosphere
Information Retrieval on the Blogosphere

Rodrygo L.T. Santos, Craig Macdonald, Richard McCreadie, Iadh Ounis

University of Glasgow
UK
{rodrygo,craigm,richardm,ounis}@dcs.gla.ac.uk

Ian Soboroff

National Institute of Standards and Technology
USA
ian.soboroff@nist.gov

now
the essence of knowledge

Boston – Delft
Foundations and Trends® in 
Information Retrieval
Volume 6 Issue 1, 2012

Editorial Board

Editor-in-Chief:
Douglas W. Oard
University of Maryland
oard@umd.edu

Mark Sanderson
RMIT University
mark.sanderson@rmit.edu.au

Editors
Alan Smeaton (Dublin City University)
Bruce Croft (University of Massachusetts, Amherst)
Charles L.A. Clarke (University of Waterloo)
Fabrizio Sebastiani (Consiglio Nazionale delle Ricerche)
Ian Ruthven (University of Strathclyde, Glasgow)
James Allan (University of Massachusetts, Amherst)
Jamie Callan (Carnegie Mellon University)
Jian-Yun Nie (Universit de Montreal)
Justin Zobel (University of Melbourne)
Maarten de Rijke (University of Amsterdam)
Norbert Fuhr (University of Duisburg-Essen)
Soumen Chakrabarti (Indian Institute of Technology)
Susan Dumais (Microsoft Research)
Tat-Seng Chua (National University of Singapore)
William W. Cohen (Carnegie Mellon University)

Full text available at: http://dx.doi.org/10.1561/1500000026
Editorial Scope

Foundations and Trends® in Information Retrieval will publish survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians
Foundations and Trends® in Information Retrieval, 2012, Volume 6, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.
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.
Contents

1 Introduction 1
   1.1 Social Media 1
   1.2 What is a Blog? 2
   1.3 Why Do People Blog? 4
   1.4 The Blogosphere 5
   1.5 Search on the Blogosphere 7
   1.6 Scope of this Survey 8

2 Blog Information Retrieval 11
   2.1 Blog Directories 11
   2.2 Existing Blog Search Engines 13
   2.3 Information Needs on the Blogosphere 17

3 Blog Post Search 21
   3.1 Ad hoc Search 21
   3.2 Opinion Search 30
   3.3 Summary 48

4 Blog Search 51
   4.1 Topical Relevance 52
   4.2 Relevance Feedback 59
   4.3 Temporal Relevance 60

Full text available at: http://dx.doi.org/10.1561/1500000026
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
1.2 What is a 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.

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](http://blogs.guardian.co.uk) or [http://www.bbc.co.uk/blogs/].
Introduction

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. 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,

---

5In 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 cre-
ating 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 cre-
ated 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 emerg-
ing 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]).
1.6 Scope of this Survey

Table 1.1. Notations used in this survey.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q$</td>
<td>A user query</td>
</tr>
<tr>
<td>$t$</td>
<td>A unigram (e.g., a term or a term feature)</td>
</tr>
<tr>
<td>$v$</td>
<td>An $n$-gram (e.g., a compound, passage or sentence)</td>
</tr>
<tr>
<td>$p$</td>
<td>A blog post</td>
</tr>
<tr>
<td>$b$</td>
<td>A blog</td>
</tr>
<tr>
<td>$d$</td>
<td>A day of interest</td>
</tr>
<tr>
<td>$s$</td>
<td>A news story</td>
</tr>
<tr>
<td>$C$</td>
<td>A corpus of items (e.g., blog posts, blogs, news stories)</td>
</tr>
<tr>
<td>$\mathcal{L}$</td>
<td>A lexicon of terms</td>
</tr>
<tr>
<td>$Q$</td>
<td>A set of queries</td>
</tr>
<tr>
<td>$D$</td>
<td>A set of retrieved items</td>
</tr>
<tr>
<td>$F$</td>
<td>A set of feedback items</td>
</tr>
<tr>
<td>$R$</td>
<td>A set of relevant items</td>
</tr>
<tr>
<td>$\mathcal{O}$</td>
<td>A set of relevant and opinionated items</td>
</tr>
</tbody>
</table>

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.
References


References


References


References


References


References


References


References


 References  121


References


124  References


References


References


