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

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is R. L. T. Santos, C. Macdonald, R. McCreadie, I. Ounis and I. Soboroff, Information Retrieval on the Blogosphere, Foundation and Trends® in Information Retrieval, vol 6, no 1, pp 1–125, 2012

ISBN: 978-1-60198-568-2

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

Editorial Board

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.

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.

Contents

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.

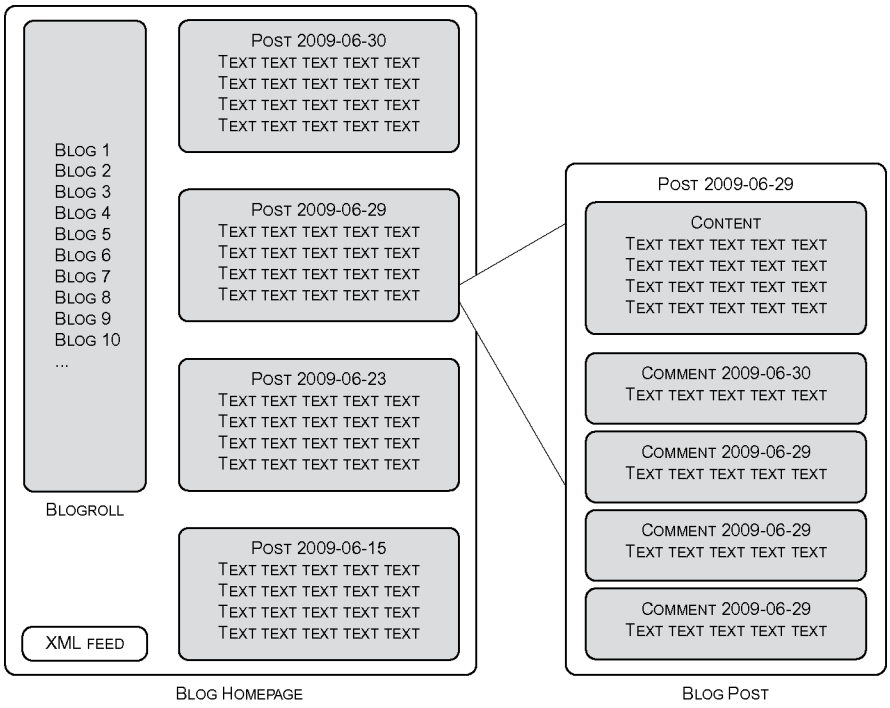

Fig. 1.1 Schematic view of a typical blog.

- 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.

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.